Statistical Topology of Reticulate Evolution

Kevin Joseph Emmett

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

COLUMBIA UNIVERSITY

2016

ABSTRACT

Statistical Topology of Reticulate Evolution

Kevin Joseph Emmett

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

List of Figures

List of Tables

Chapter 1

Introduction

Darwin's *On the Origin of Species* contains a single figure, depicting the ancestry of species as a branching genealogical tree [22] (see Figure 1.1). Since then, the tree structure has been the dominant framework to understand, visualize, and communicate discoveries about evolution. Indeed, a primary focus of evolutionary biology has been to expand the *universal tree of life*, the set of evolutionary relationships among all extant and extinct organisms on Earth [8]. Traditionally, this was the realm of phenotype-derived taxonomies. With the advent of molecular sequene data and computational approaches for tree-inference, molecular phylogenetics has become the standard tool for inferring evolutionary relationships. However, a tree is accurate only if the Darwinian model of descent via reproduction with modification is the sole process driving evolution. It has long been recognized that there exist alterative evolutionary processes that can allow organisms to exchange genetic material through means beyond reproduction [3]. Notable examples include horizontal gene transfer in bacteria, species hybridization in plants, and meiotic recombination in eukaryotes. Collectively, these processes are known as *reticulate evolution*. These stand in contast to descent with modification, an example of *clonal evolution*. [1] Increasing genomic data, powered by new high-throughput sequencing technologies, has shown that these reticulate processes are more

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

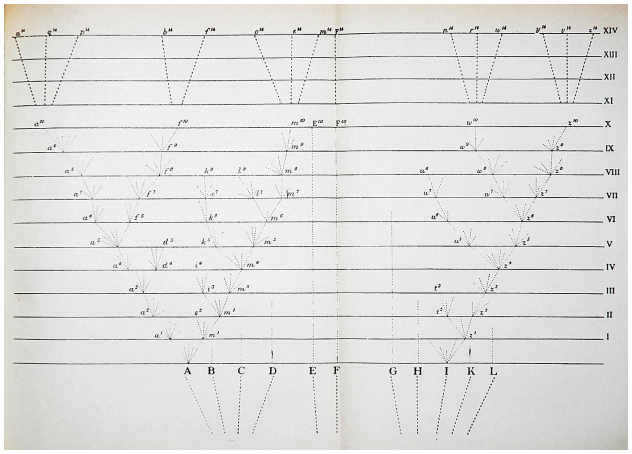


Figure 1.1: The only figure in Darwin's *On the Origin of Species*. Darwin argued for descent with modification and natural selection as the driving processes underscoring evolution. In this figure, Darwin sketched his idea for how diverging species would result in a tree structure. Reproduced from [22].

prevalent than originally expected. For some, this has called into question the tree of life hypothesis as an organizing principle and prompted the search for new ways of representing evolutionary relationships [24, 57].

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

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

1.1 Molecular Evolution and the Tree Paradigm

The combination of Darwin's theory of natural selection with Mendelian genetics led to the *modern evolutionary synthesis*, outlined in the first half of the twentieth century in pioneering works by Ronald Fisher, Sewall Wright, JBS Haldane, and others. [2] The modern synthesis was based largely on an analysis of distributions of allele frequencies in distinct populations, the purview of classical population genetics. The field was placed on a molecular foundation with Watson and Crick's discovery of the DNA double-helix in 1953 [64]. These developments led to the establishment of *molecular evolution*, the analysis of how processes such as mutation, drift, and recombination act to induce changes in populations and species.

The information underlying an organism's form and function is encoded in its genome, the complete sequence of DNA contained in each cell. The genome can be represented as a string of nucleotides, indexed by position. Embedded within the genome are regions defining the genes which code for functional proteins, as well as non-coding regions which have as-yet unknown function. [3] When an organism reproduces, either sexually or asexually, a complete copy of the genomic information is passed to the offspring. Because the molecular mechanisms that control this copying are not exact, errors in replication are introduced. These errors can take the form of single point mutations (or single nucleotide polymorphisms, SNPs) or small insertions and deletions of a few nucleotides (indels). [4] Under the neutral theory of evolution, the majority of these errors will have very little impact, either positive or negative, on the descendant organism. A small fraction of mutations will result in an appreciable fitness differential compared to other organisms, and it is on these organisms that natural selection will act.

---

[2] See [41] and [37] for historical detail.

[3] In humans, only 1.5% of the genome is protein-coding, the rest largely non-functional. [45]. Up to 5-8% of the human genome is believed to consist of endogeneous retroviruses, dead viruses which have integrated their genome into the human genome.

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

While molecular biology has largely focused on the biochemical and biophysical mechanisms underlying these processes, *molecular phylogenetics* has focused on the comparative analysis of macromolecular sequences to infer genealogical and evolutionary relationships. Molecular phylogenetics began with Emile Zuckerkandl and Linus Pauling's recognition in the early 1960's that the information encoded in a set of molecular sequences could itself be used as a document of evolutionary history [69, 70]. It became apparent that given two sequenced organisms, counting the differences between their respective sequences could be used as a quantitative measure of the amount of evolutionary divergence between the two. If one has a larger set of sequenced organisms, computing the complete set of pairwise distances gives a *distance matrix* for the organisms. From the distance matrix, one attempts to associate a tree to the data such that pairwise distances along the tree are close to the pairwise sequence distances. Walter Fitch and Emanuel Margolish helped popularized this approach by constructing a weighted least squares approach to fitting phlyogenetic trees from distance matrices [32]. Since that time, the development of numerical approaches for inferring evolutionary relationships has evolved into a mature discipline and the use of molecular sequence data to infer phylogeny has become a standard practice across a wide range of biology and ecology. While other approaches to tree inference have been developed, including parsimony, quartet analysis, and Bayesian methods, we will focus on distance matrix methods because of their close connection with the topological ideas we employ later.

One important early result from molecular phylogenetics was Carl Woese's organization of bacteria, eukarya, and archaea into the three domains of life [66]. Prior to Woese, there were two recognized domains of life: prokaryotes, single-celled organisms lacking a nucleus, and eukaryotes, multi-celled organisms with an enveloped nucleus. Using 16S subunit ribosomal RNA sequencing, Woese discovered that the prokaryotic domain actually split into two evolutionarily distinct groups. One of these, which he termed *archaebacteria* was more closely related to eukaryotes than were there the rest of the prokaryotes. This led to the three-domain system of life (see Figure 1.2).

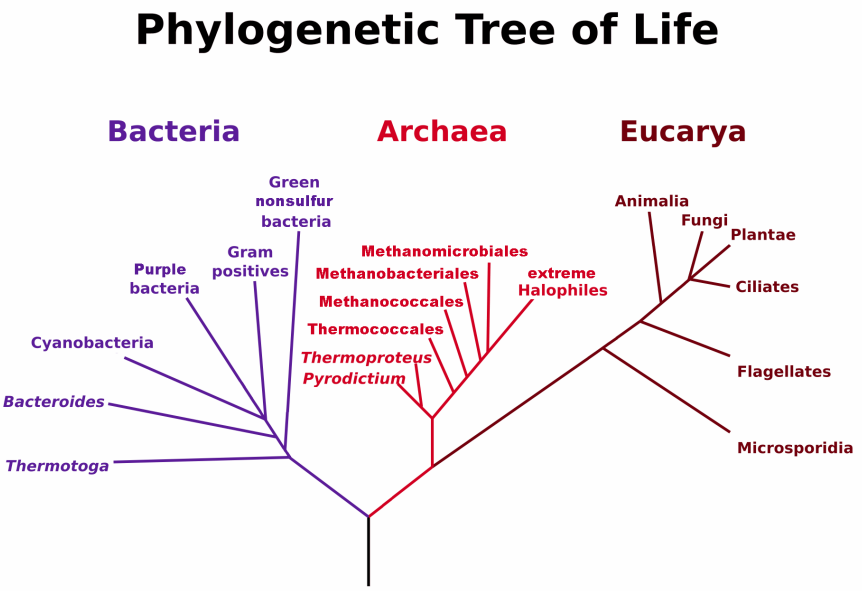


Figure 1.2: Carl Woese's three kingdom tree of life. Using 16S subunit ribosomal RNA, Carl Woese identified archaea as a distinct phylogenetic kingdom. Previously, based on morphological similarity (specifically, unicellular and lacking a nucleus), archaea had been grouped with bacteria. This result was an early success for molecular phylogenetics and the use of conserved gene segments for molecular classification. Figure adapted from [67].

This work had several important consequences. First, it established the use of molecular data to inform about large-scale patterns of evolutionary history. Using only morphological data had led to an inconsistent classification of archaea. Second, it positioned 16S rRNA profiling as the primary source of data for use in comparative genomics. The use of this genomic region was justified on the basis of being one of the few universal gene segments that is conserved across all species. Constructing a universal tree is predicated on there being orthologous genes, shared genes related through speciation events, that can provide a common foundation for comparative study. Finally, it solidified the tree paradigm as an organizing principle for relating extant species. Even though reticulate processes had been known since the nineteenth century, the idea that evolutionary relationships should be described by a bifurcating tree had been paramount since Darwin. Reticulate processes were either ignored completely, or expected to occur at such low frequencies that they need not be considered.

1.2 Reticulate Processes and the Universal Tree

Despite the significant impact of Woese's observation, there remains a subtle difficulty, which Woese himself would come to contemplate in later work [65, 36]. 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 [21]). Even more striking, this accounts for less than 0.00005% of the human genome. While recent work has developed approaches for constructing reference trees from larger gene sets [19], the fact remains that the vast majority of genomic information is *not* incorporated into the tree.

The reason for this situation is twofold. First, not all genes are shared universally across all species. In constructing a phylogenetic tree using sequence data, only genes that are present across all species are informative. Second, even among universal genes, the presence of reticulate evolutionary processes will confound systematic analysis. The model of a bifurcating tree will be consistent only if all loci share the same pattern of bifurcation. When organisms exchange genetic material by means other than direct reproduction, the ancestral relationships between species will depend on which genomic regions are used. If one were to use two different genomic regions, two different tree topologies may be generated, with conflicting phylogenetic relationships. It remains an open question how to best construct a consistent evolutionary history from conflicting phylogenetic signals[5]

Historically, reticulate processes were believed to occur at such a low frequency that they could be safely ignored when considering evolutionary relationships. However, new genomic data has shown that, particularly in microorganisms such as bacteria and archaea, reticulate processes are much more prevalent than originally expected [56]. Incompatibilities in the tree paradigm now appear as the rule, not the exception, which has led to calls for new representations of evolutionary relationships [24, 25]. Many have argued that, in light of new genomic evidence, the very notion of a universal tree of life must be discarded [43, 44].

[5]There exists a cottage industry of methods for aggregating conflicting *gene trees* into a consensus *species tree*, see [48].

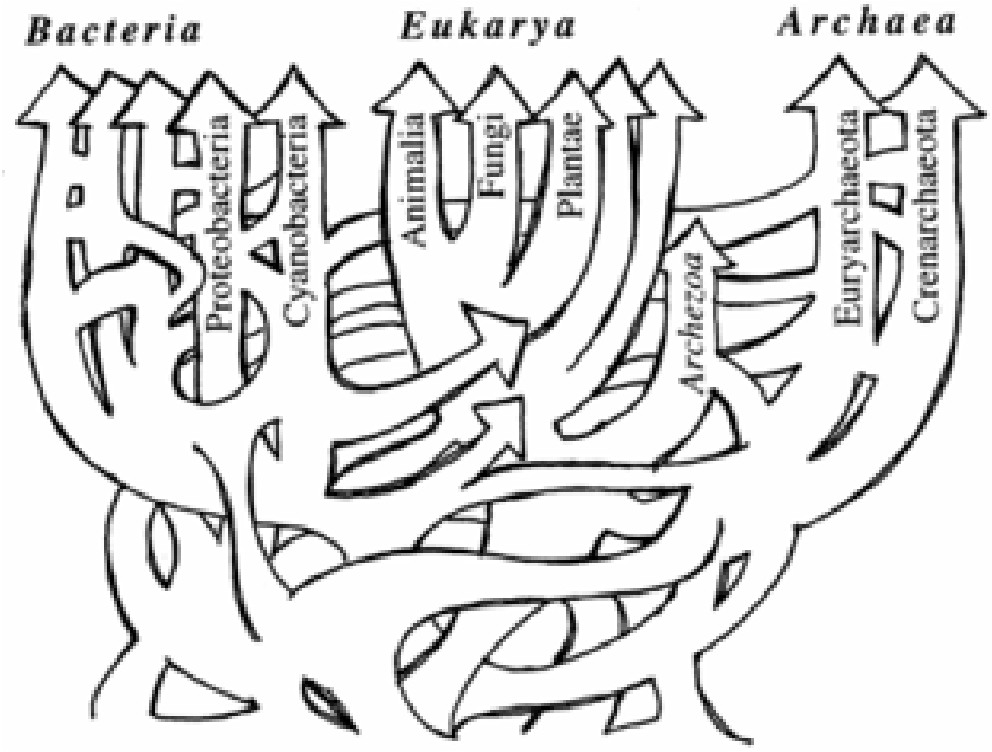


Figure 1.3: W Ford Doolittle's representation of the universal tree of life with reticulate evolution. While the three domains of life are still recognizable, patterns of divergence no longer follow a strictly treelike model. (From *Science*, vol. 284, issue 5423, page 2127. Reprinted with permission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

Finally, reticulate evolutionary processes are of more than just theoretical concern. In HIV, frequent homologous recombination confounds our understanding of the epidemic's early and present history [12]. In influenza, segmental gene reassortments lead to antigenic novelty and the emergenence of epidemics [54]. In several pathogenic bacteria, including *E. coli* and *S. aureus*, horizontal gene transfer has been responsibile for the spread of antibiotic resistance genes [1, 23].

1.3 Evolution as a Topological Space

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

Topology as a mathematical field is concerned with properties of spaces that are invariant

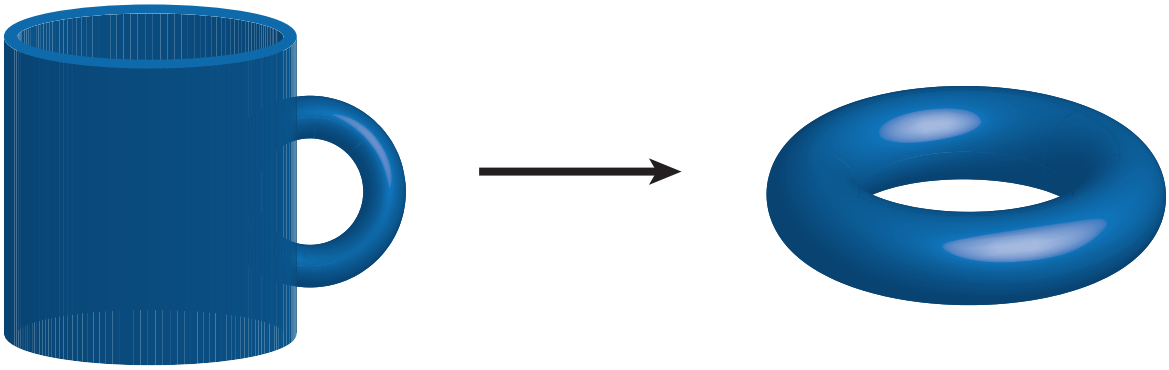


Figure 1.4: The paradigmatic example of topological equivalence. The coffee mug can be continuously deformed into the donut and are therefore topologically equivalent. Both exhibit the topology of a solid torus ($D^2 \times S^1$).

under continuous deformation. Such properties can include, for example, connectedness and the presence of holes. Two objects are considered topologically equivalent if they can be deformed into one another without introducing any cuts or tears. As a paradigmatic example, consider the coffee mug and the donut (Figure 1.4). While seemingly different, it is not difficult to see that both objects consist of a single connected component that is wrapped around a single hole. Were the objects smoothly pliable they could be freely deformed into one another. Topologically, the two objects are equivalent.[6]

Algebraic topology quantifies our intuitive notions of shape by associating algebraic structures to different invariants. For our purposes, the most relevant invariants will be the *Betti numbers*. We give a more complete characterization of Betti numbers in Chapter 2, but the intuition is as follows. The Betti numbers are a collection of integers indexed by integer n describing the connectivity of a space at different dimensions. First, we can think of b_0 as representing the number of connected components, or clusters, in our space. Next, we can

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

think of b_1 as representing the number of one-dimensional loops in our space. Equivalently, this is the number of cuts needed to transform the space into something simply connected. Higher Betti numbers, b_n for $n > 1$ will correspond to higher dimensional holes. In our coffee mug example, because both objects have the same Betti numbers ($b_0 = 1$, $b_1 = 1$, and $b_n = 0$ for $n > 1$), they can be considered topologically equivalent. Our goal in this work will be to adopt a similar perspective as this example and characterize evolutionary spaces as topological spaces using their Betti numbers.

To give the very simplest example, consider Figure 1.5. The example presents two possible scenarios describing the evolutionary relationships of three species, labeled a, b, and c. The objects are to be read such that moving vertically corresponds to moving backwards in time. Branch lengths will correspond to some notion of evolutionary divergence. Internal vertices represent extinct ancestors of the three species, up to the root of the tree, r, which represents the most recent common ancestor. On the left, we have a simple tree topology relating the three species. Considering the shape of the tree, there is a single connected component, giving $b_0 = 1$. Further, we see that there are no loops formed by the branches, giving $b_1 = 0$. The object is trivially contractible, a property which will hold for all tree topologies. On the right, we have a reticulate topology relating the three species. We can envision species b as being the reticulate offspring of parents ancestral to species a and c. That is, species b carries unique genetic material from both species a and species c. To account for this, two branches merge into the vertex that is directly ancestral to b. Considering the shape, there is again a single connected component, giving $b_0 = 1$. However, because of the reticulate event mixing material from a and c, there is now a loop formed in the topology, giving $b_1 = 1$ The object is no longer treelike and is characterized by a nontrivial topology. The Betti numbers capture the essential difference in the two evolutionary histories.

Consider again Darwin's branching phylogeny (Figure 1.1) and Doolittle's modified tree accounting for reticulate evolution (Figure 1.3). The two objects can be imagined to be representations of two different topological spaces. Darwin's branching phylogeny is a tree

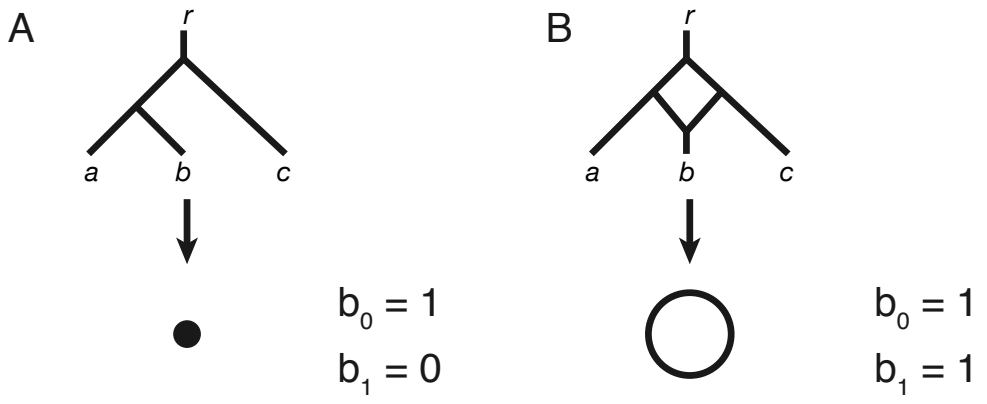


Figure 1.5: (A) A simple treelike phylogeny is contractible to a point. (B) A reticulate phylogeny that is equivalent to a circle and not contractible without a cut. The two spaces are not topologically equivalent and can be distinguished by their Betti numbers.

and hence trivially contractible ($b_n = 0$ for $n > 0$). In contrast, Doolittle's construction has a much more complex topology, with loops being formed at points where reticulate events have occurred. The object will have nonvanishing Betti numbers, which will be associated with the amount of reticulation. The remainder of this thesis focuses on expanding this idea and applying it to real data sets with the goal of measuring the prevalence and scale of reticulate evolutionary events. Our aim will be to characterize reticulate exchanges of genetic material by the parental sequences involved in the exchange, by the amount and identity of material exchanged (i.e., the genes or loci involved), and the frequency with which similar exchanges occur. Several important questions will be dealt with, such as how to construct topological spaces from finite samples, how to make comparisons among gene sets, and how to make statistical statements about reticulate events. We will address these questions, and in doing so develop new techniques to construct and extract topological and statistical information from evolutionary data. In doing so, we provide a fuller understanding of evolutionary relationships than allowed by current phylogenetic methods.

1.4 Thesis Organization

The remainder of this thesis is organized as follows.

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

In Part **??**, we develop two complementary approaches for analyzing sequence data using TDA. In Chapter **??**, we propose methods of constructing topological spaces that generalize standard constructions but are suite to the particular requirements of phylogenetic applications. We draw on previous work in phylogenetic networks and use homology to provide a quantitative assessment of reticulate processes. This work was published in [27]. In Chapter **??**, we develop methods for performing statistical inference using summary statistics computed using methods from TDA. This is the first such use of TDA as a tool for performing parametric inference and should generalize to a wide range of application settings. This work was published in [28]

In Part **??**, we apply our approach to several problems in evolution and genomics. In Chapter **??** we study bacteriophages. In Chapter **??** we study influenza. In Chapter **??** we study pathogenic bacteria and use topological techniques to represent the spread of antibiotic resistance. In Chapter **??** we study prokaryotic evolution and species tree topologies. In Chapter **??** 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 **??** we analyze Hi-C data to explore patterns of chromatin folding in the nucleus in both prokaryotic and human datasets.

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

Chapter 2

Background

This thesis uses newly developed approaches from applied topology to study problems in evolutionary biology and genomics. In this chapter we provide background material on the In Section 2.1 In Section 2.2 Finally, in Section 2.3

Here we supply sufficient background to motivate our approach.

2.1 Biology

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

2.1.1 Genes and Genomes

The information required to express an organism's biological form and function is contained in the genome. Physically, the genome is manifest as a sequence of nucleotides (DNA), at least one copy of which is packaged inside each cell of an organism. Abstractly, the genome

is represented as a linear sequence of characters defined over the alphabet $\{A, C, G, T\}$.[1] Contained in this sequence are subsequences representing genes, which code for the protein products that ultimately affect function. Further embedded in the genome is a complex regulatory pattern of transcription factors controlling the expression of particular genes and directing cellular differentiation and development.

Following the central dogma of biology, DNA is transcribed into RNA, RNA is translated into amino acids, and amino acids are folded into proteins [20]. 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 bserved diversity. The field concerned with exploring these relationships is *comparative genomics*.

2.1.2 Evolutionary Processes

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

2.1.2.1 Clonal Evolution

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

It is clonal evolution that Darwin had in mind when he described the idea of descent with modification, whereby a parent passes genomic information to an offspring subject to

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

random drift. Importantly, because there is always a direct parent–offspring relationship, clonal evolution will be consistent with a phylogenetic tree model.

2.1.2.2 Reticulate Evolution

Reticulate evolution, or horizontal evolution, refers to exchange or acquisition of genetic material via processes that do not reflect a direct parent–offspring relationship. As we will see, these processes can make inferences about historical evolutionary relationships difficult. Different types reticulate processes occur in different types of organisms (summarized in Table 2.1.2.2).

Viruses replicate by infecting a host cell and then using the host cell machinery and resources to produce multiple copies of viral genetic material. The genetic material is then packaged into new virus particles which are shed off in order to infect new cells. Reticulation can occur when two virus particles coinfect the same host cell. During the replication process, genetic material can be exchanged in one of two ways: *reassortment* or *recombination* (the two processes are contrased in Figure 2.1). Reassortment occurs in viruses whose genomes are segmented, such as influenza. Segments are similar to chromosomes, such that a single virus particle will contain a single copy of each segment. Reassortment occurs when coinfection results in packaging of segments taken from different virus particles. The result viral progeny will then be a genetic mixture of segments from each parental strain. Recombination, more common in nonsegmented viruses such as HIV, involves a break-rejoin mechanism during the replication process. Here, an error in the polymerase during replication can result in an incomplete copy of the genome (a break). At this point, several cellular processes involved in repair can be recruited to complete the replication process using a homologous region. If coinfection has occurred, it is possible for these processes to initiate repair using material from a different parental strain. The outcome will be novel genetic material that includes a crossover from one strain to another. Break-rejoin crossover is a type of *homologous recombination*.

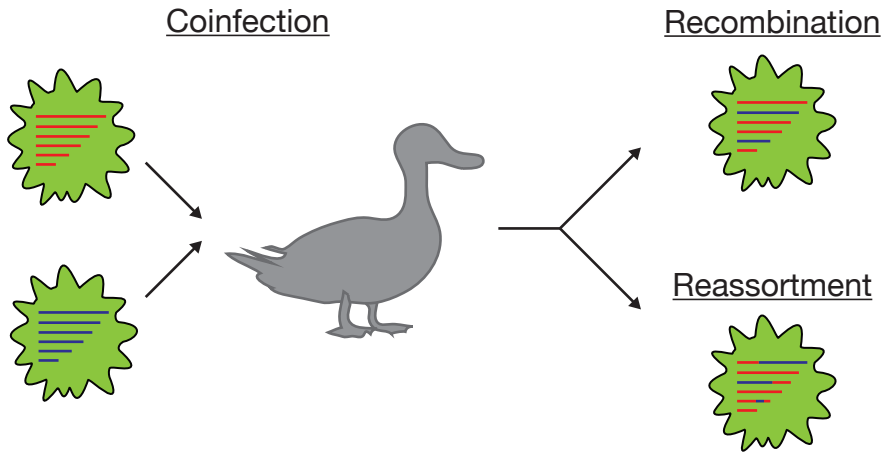


Figure 2.1: The two modes of viral reticulation. Coinfection of the same host cell can lead to either reassortment, in which whole viral segments are exchanged, or recombination, in which breakpoints can occur within segments. The former process is common in influenza, the latter in HIV. The end result,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

In bacteria and other prokaryotes, reticulate evolution can occur when foreign DNA from a donor is acquired by a target organism and integrated into its genome. Three generic mechanisms have been identified, depending on the route by which foreign DNA is acquired [56]:

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

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

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

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

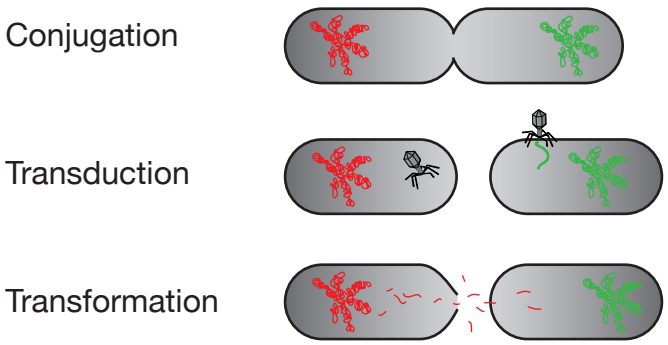


Figure 2.2: Three modes of viral reticulation. (1) Conjugation, in which direct cell-to-cell contact results in transfer of genetic material; (2) Transformation, in which foreign DNA is acquired via uptake from freely circulating DNA in the environment; and (3) Transduction, in which exchange of genetic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

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

First, hybrid speciation refers to the cross-breeding of animals or plants of different species. This mixing of genetic material can lead to the development of a third species with a phenotype distinct from both parents. Hybrid speciation was originally believed to be a rare occurrence in nature and hybrid offspring to be infertile. However, recent genomic data has demonstrated that hybrid speciation occurs quite frequently in plants [2, 3]. Indeed, Mendel's early experiments in hybridization were themselves an artifically induced form of reticulate evolution.

Second, meiotic recombination refers to a specialized process for generating diversity that occurs in sexually-reproducing polyploid organisms, such as humans, during meiosis. Meiosis is the process by which a single cell containing n copies of each chromosome results in four distinct cells each with $n/2$ copies of each chromosome. These special cells are called gametes. Sexual reproduction consists of the fusion of two gametes during fertilization to form a zygote, which ultimately develops into a viable offspring. Meiosis is a multistep process consisting of an initial round of DNA replication followed by two rounds of cell

division. Meiotic recombination occurs after the initial round of DNA replication and prior to cell division. After DNA replication, there are two copies of each homologous chromosome that are joined at a centromere. The two sets of chromosomes then pair with each other and exchange DNA through physical interactions known as crossovers.[2] This is another example of homologous recombination and results in new allelic patterns mixing genetic information from both parents.[3] After crossover occurs, two phases of cellular division result in gametes with $n/2$ copies of each chromosome.

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

[2]These crossovers have been shown to not occur randomly, but rather at recombination hotspots regulated by binding motifs for by the PRDM9 protein. Cite Pablo Paper and a Molly paper.

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

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

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

However, in the case where there is substantial disagreement among gene trees, the very notion of a species tree may be flawed. Traditionally, evolutionary biology has concerned itself with characterizing relationships in light of vertical evolution alone. However, increasing evidence (what evidence?) has pointed to the important role played by horizontal evolution, particularly in prokaryotic evolution. [Further citations and exposition of Doolittle, Koonin, and Gogarten.]

2.1.3 Mathematical Models of Evolution

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

netics models was comparative studies of allele frequencies across populations. These studies have primarily been replaced by large-scale genomic surveys which have provided unprecedented insight into ancient population structure and historical migrations.

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

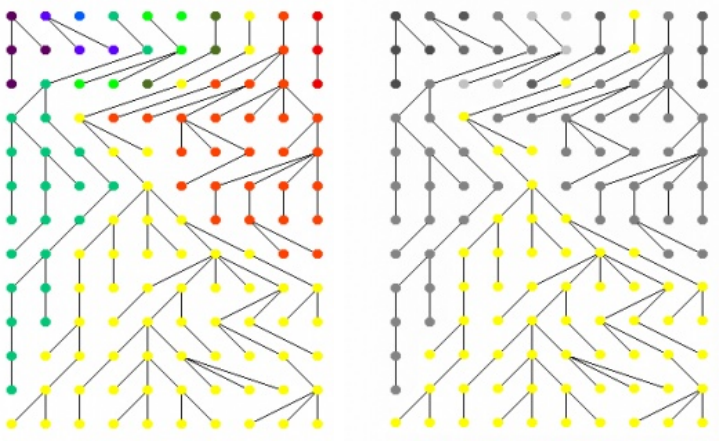
2.1.3.1 The Wright-Fisher Model

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.

2.1.3.2 The Coalescent Process

The coalescent process is a stochastic model that generates the genealogy of individuals sampled from an evolving population [63]. 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:

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



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

Figure 2.3: Two models for simulating evolutionary data.

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` [39].

2.1.3.3 Metrics on Sequences

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

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

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

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

where p is the proporation of sites that are different.

2.1.4 Phylogenetic Methods

A phylogenetic tree is a binary tree in which leaves are associated with particular species or taxa, and the branching pattern of the tree reflects diverging evolutionary relationships. Branch lengths on the tree are associated with evolutionary divergence between sets of taxa. Molecular phylogenetics refers to a large collection of methods for inferring branching pat-

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

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

terns from aligned molecular sequence data.[6] In general, the problem of finding an optimal tree associated with sequence data is NP-complete [33], however several approximate methods have been developed. The primary types of methods include maximum parsimony, distance-matrix methods, maximum likelihood (ML), and Bayesian inference. Maximum parsimony attempts to find the phylogenetic tree that minimizes the number of evolutionary changes required to explain the observed sequences. Distance-matrix methods first compute a matrix of pairwise distances between taxa and then find the tree that best approximates these distances. ML and Bayesian methods use specific models of evolution to assign probability distributions over trees. In this work we concentrate on distance-matrix methods because of their close connection with the finite metric spaces considered in applied topology.

2.1.4.1 Distance-Matrix Methods

Given a set of aligned molecular sequences, distance-matrix methods first compute the pairwise matrix of genetic distances using one of the metrics as described in Section 2.1.3.3. Then, the binary tree that best approximates those distances is iteratively fit to this data. This approach to phylogenetic inference were introduced by Cavalli-Sforza and Edwards in 1967 [16] and Fitch and Margoliash in 1967 [32]. The Fitch-Margoliash method uses a weighted least squares approach to tree-fitting, such that larger distances are weighted less, due to higher chances for random error. Distance-matrix methods are popular for their high speed and scalability as well as high accuracy in most cases.

Currently, the most widely implemented distance-matrix method is neighbor-joining.[7] One particular reason neighbor-joining is popular is that under certain conditions, discussed below, it has been shown to exactly recover the correct tree. The neighbor-joining algorithm is described in Algorithm 1.

[6] See Felsenstein, *Inferring Phylogenies* for a readable and thorough introduction to the field [31].

[7] Neighbor joining was introduced by Saitou and Nei in 1987 [58].

Data: $n \times n$ distance matrix D
Result: Phylogenetic tree on n leaves
while *rule* **do**
  Compute Q matrix;
end

Algorithm 1: The Neighbor Joining Algorithm. Adapted from Wikipedia entry on Neighbor-Joining.

2.1.4.2 Additive Metrics and the Four Point Condition

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

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

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

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

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

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

Sequence data can fail to be additive for several reasons. First, sequencing error. Errors can introduce noise into the measured genetic distances. Second, homoplasy. A homoplasy occurs when the same mutation is introduced multiple times in a set of organisms. The presence of homoplasy will underestimate genetic distance between taxa. Third, reticulate evolution. As described previously, in cases of reticulate evolution no tree will accurately describe the observed data.

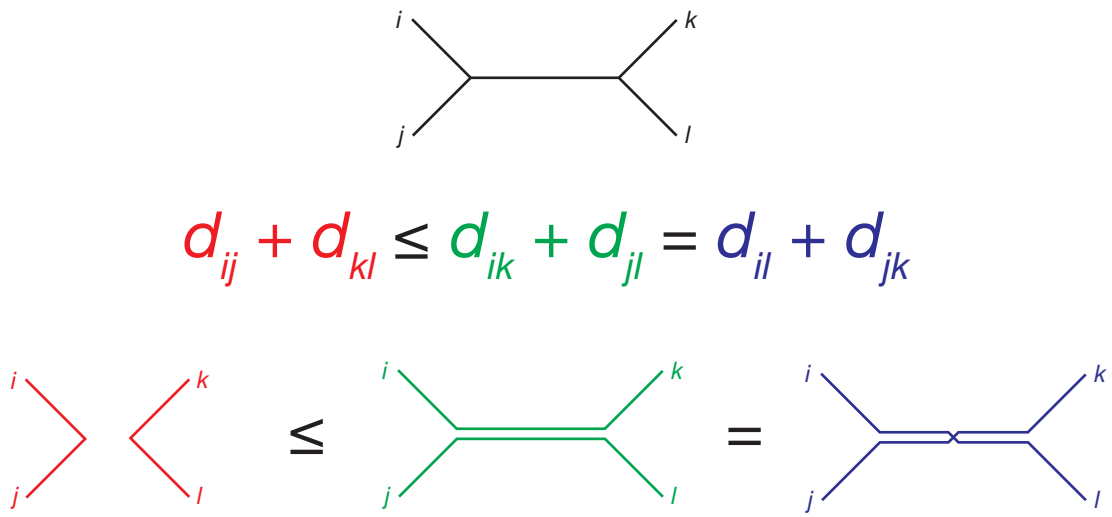


Figure 2.4: A visual interpretation of the four point condition for additivity. For any four leaves, there exists a labeling $\{i, j, k, l\}$ such that $d_{ij} + d_{kl} \leq d_{ik} + d_{il} = d_{il} + d_{jk}$. Of the three possible ways of arranging the sums of distances, two will involve traversing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

2.1.4.3 Number of Tree Topologies

The number of unrooted bifurcating tree topologies with L leaves is $(2L-5)!!$.[8] This can be easily shown using induction. For $L = 3$, we have $\mathcal{T}(3) = 1$ and 3 branches. To pass to $L = 4$, we can add the fourth leaf to any of the 3 branches, resulting in 3 different topologies. For $L = 4$, we have $\mathcal{T}(4) = 3$. Every time we add a leaf, we add two branches – one external and one internal. For $L = n$, we have $\mathcal{T}(n) = (2n-5)!!$ and $2n-3$ branches. For $L = n+1$, we can add the new external branch to any of the current $2n-3$ branches. A rooted tree with L leaves can be considered as an unrooted tree with $L+1$ leaves. Therefore, the number of rooted bifurcating tree topologies with L leaves is $(2L-3)!!$ As can be seen, the number of tree topologies explodes with the number of leaves.[9]

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

[9] It was observed by Walter Fitch that for more than 20 species there are more than Avogadro's number of topologies.

2.1.4.4 The Space of Phylogenetic Trees

A phylogenetic tree with L leaves is characterized by its topology and the lengths of each branch. As shown in the previous section, there are $(2L-5)!!$ possible unrooted topologies. There are $2L-3$ branches: L external branches and $L-3$ internal branches. Tree space refers to an abstract construction for representing each possible tree as a point in a geometric space. Studies of tree space were initiated by Billlera, Holmes, and Vogtmann (BHV) in [7]. In that paper, the authors studied rooted trees with zero-length external branches, a space we denote as BHV_L. A geodesic distance was defined between trees of different topology. This analysis was extended by Zairis *et al.* in [68], in which unrooted trees with non-zero external branches were considered. The external branches are constrained to sit in the positive open orthan $(\mathbb{R}^{\geq 0})^L$. An evolutionary moduli space is then defined as the product

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

The tree space construction allows one to define statistics, such as means and variances, on collections of trees in a meaningful way. An interesting statement that can be made within this framework is that phylogenetics is essentially the projection of real data onto tree space.

We show an exaple of the tree space construction on $L = 4$ and $L = 5$ leaves in Figure 2.5. The case of $L = 4$ is particularly simple to analyze. There are three topologies, corresponding to the patterns $((a,b),(c,d))$, $((a,c),(b,d))$, and $((a,d),(b,c))$. Each topology has a single internal branch. Attached to each external topology is a space $\mathbb{R}^4$ for the external branches. The case of $L = 5$ also has a relatively simple structure. There are fifteen possible topologies, each with two internal branches. The topologies are arranged as a Petersen graph, a common construction in graph theory. Vertices of the Petersen graph correspond to degenerate points on one of the two internal branches.

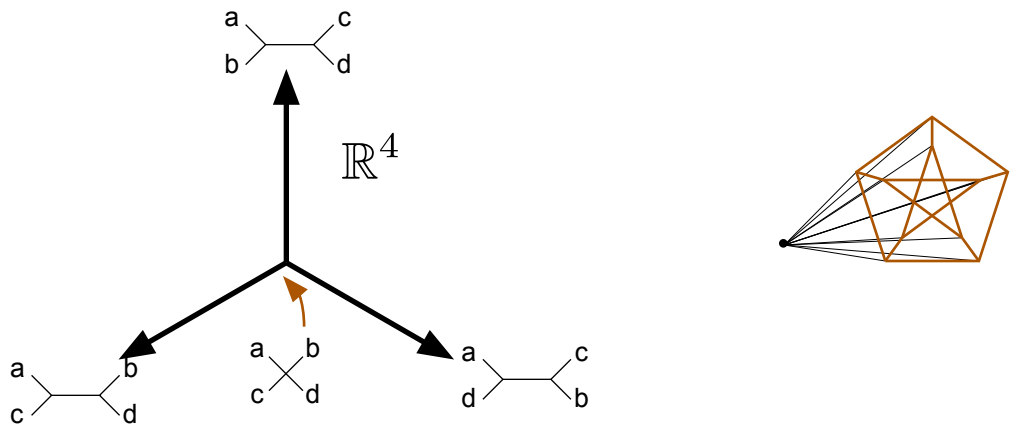


Figure 2.5: Examples of Tree Space. (A) Tree space on 4 leaves. (B) Tree space on 5 leaves.

2.1.4.5 Phylogenetic Networks

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

2.2 Topological Data Analysis

Topology is the branch of mathematics that formalizes our intuitive notions of shape. More concretely, topology provides the methods to characterize the properties of objects and spaces that remain invariant under continuous deformation. For example, squeezing a circle into an ellipse by compressing along one axis does not change the fact that the object encloses a

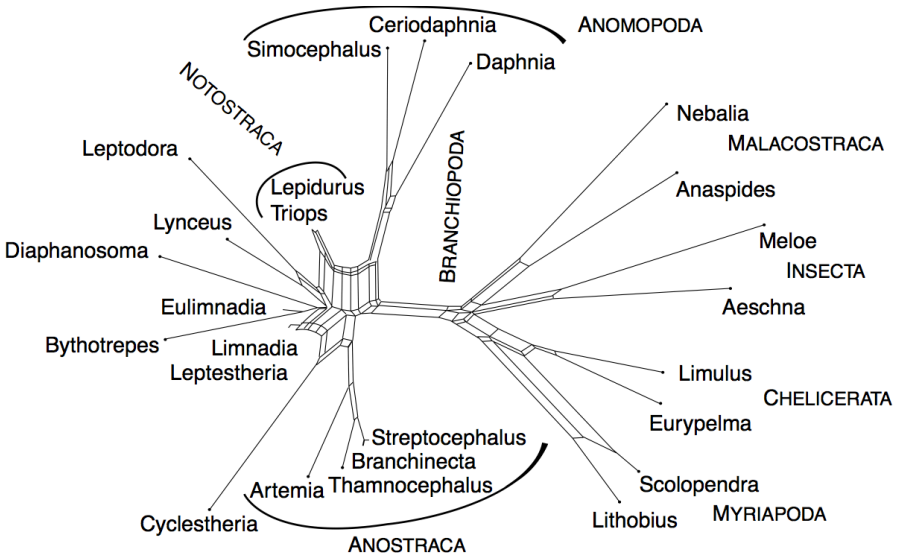


Figure 2.6: Example of a split network of genus Branchiopoda and outgroups. Computed using the Neighbor-Net algorithm. Phylogenetic incompatibilities are represented by conflicting splits. reprinted from BMC Evolutionary Biology 7:147 (2007).

single loop. Or, as we saw in the introduction, the coffee mug can be continuously deformed into the donut. Likewise, if we take a tree and change the lengths of its branches, the tree remains a tree.[10] In each of these examples, while the deformation has substantially altered *local* properties of the space, on a *global* level certain essential characteristics have remained unchanged. From the perspective of topology, the spaces are considered identical. The question then is how to formalize our idea of global shape in order to systematically reason about it.

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

While topology has traditionally developed through the study of abstract spaces, leading to very rich and beautiful constructions[11], real data does not come in the form of perfect continuous spaces. Recent effort over the past 15 years has focused on developing methods to apply topology to real world problems in science and engineering. This work, collectively falling under the heading of *topological data analysis* (TDA), has focused on efficient algorithms for computing topological invariants from finite, noisy data. TDA now encompasses a wide range of efforts and can now be considered a branch of applied mathematics in its own right. It has emerged from substantial interdisciplinary effort between mathematicians,

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

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

computer scientists, and domain experts.

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

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

The two main methods from TDA that we employ are *persistent homology* and *mapper*. Persistent homology provides a way to efficiently compute the topological invariants of a space across multiple scales, while mapper provides an approach for condensed representation and visualization of high-dimensional data. In this section, we provide an overview and discussion of these two methods from the perspective of an end-user, treating each method as a pipeline for transforming from raw data to a concise topological summary. While the mathematical literature on these methods is extremely deep, our goal is to explain things in sufficient detail for a wide audience to grasp the main ideas. We therefore include a brief introduction of the basic mathematical concepts we employ. The primary concept we require is *homology*, a particular way in which topological invariants can be assigned to spaces.

The following sections draw on several excellent reviews of TDA, including [13], [26], and

[35]. A more thorough introduction to algebraic topology can be found in [38].

2.2.1 Preliminaries

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

2.2.1.1 Simplices and Simplicial Complexes

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

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

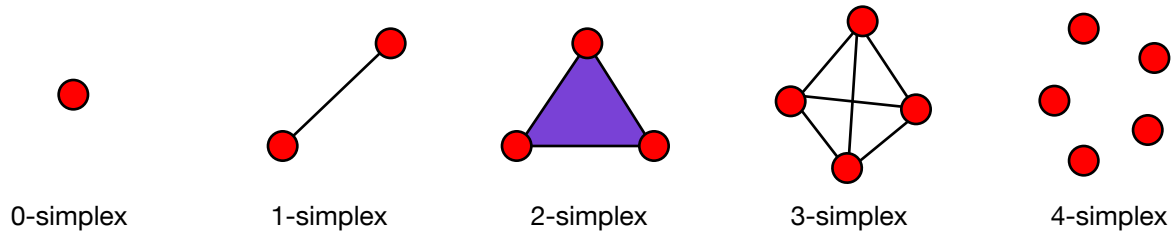


Figure 2.7: Simplices are the fundamental building blocks of our topological structures. They can be thought of as triangles generalized to arbitrary dimension. Here we show k-simplices for $k = 0$ to $k = 4$.

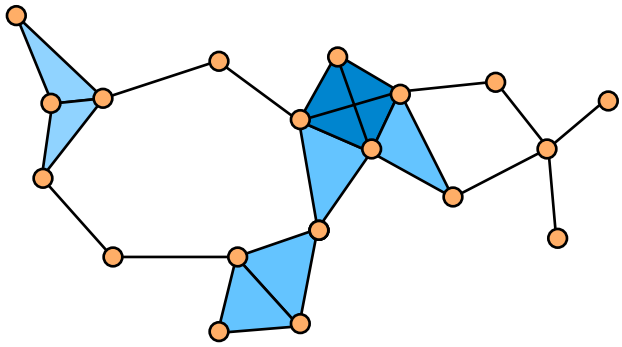


Figure 2.8: A finite simplicial complex is an object built from finite number of simplices glued together in a sensibly nice way.

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

An important operator is the boundary operator, $\partial : C_k \to C_{k-1}$. The boundary of a

[12]In general, an algebraic topology can be defined with coefficients in arbitrary fields. We use $\mathbb{Z}_2$ for simplicity, efficiency, and because properties, such as torsion, that arise over more complex fields are not expected to be present in the biological data we consider. It is important to keep this in mind, as it was in fact shown that torsion can arise in real data in [14]. In that paper, an association was shown between the space of natural images and the Klein bottle.

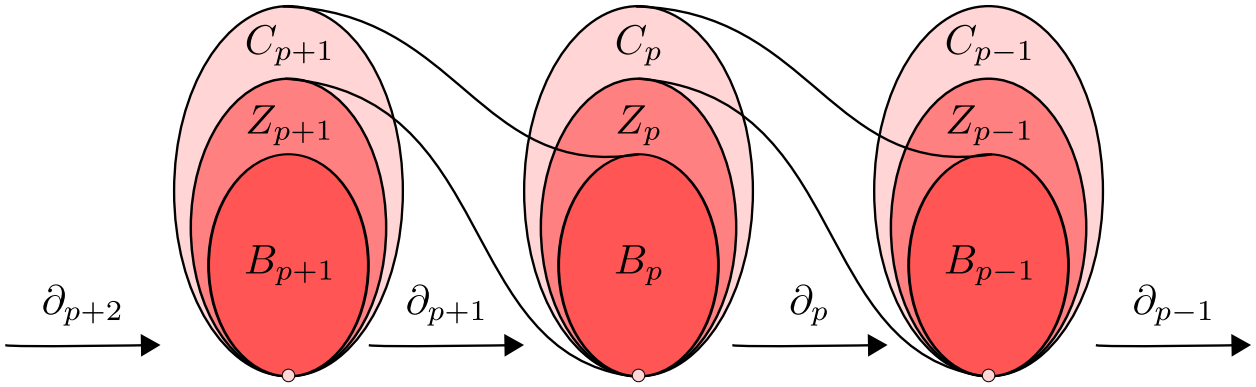


Figure 2.9: Relationship between the chain group (C_p), cycle group (Z_p), and boundary group (B_p). We show the action of the boundary operator ∂_p on each group at each dimension. Adapted from [29].

simplex σ, $\partial_k \sigma$, is the sum of its facets.

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

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

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

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

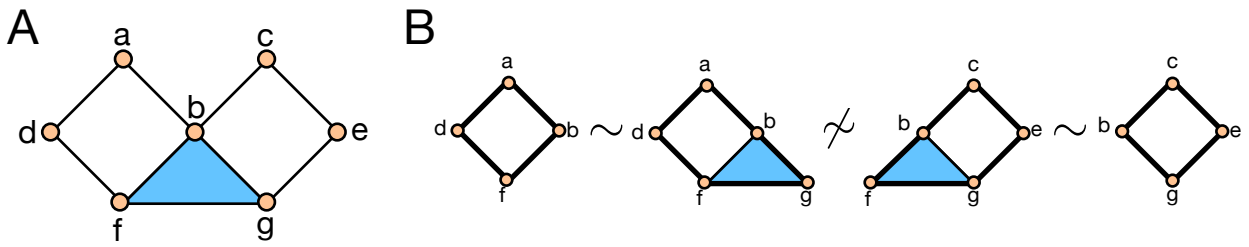


Figure 2.10: (A) A simplicial complex defined on a set of 7 vertices, $S = \{a, \ldots, g\}$. The object has one connected component ($b_0 = 1$) and two holes ($b_1 = 2$). (B) Four cycles that can be defined on the complex. Cycles $z_1 = \{(a, b) + (b, f) + (f, d) + (d, a)\}$ and $z_2 = \{(a, b) + (b,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

2.2.1.2 Homology

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

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

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

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

2.2.1.3 Constructing Complexes From Data

Finally, we must consider how to construct a simplicial complex from a given point cloud S.[13] There are two common constructions we will describe: the *Čech complex* and the *Vietoris-*

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

Rips complex. Both constructions involve a scale parameter ϵ, and balls of radius ϵ placed at the center of each vertex in S. Edges are drawn between vertices when balls overlap, that is, when $d(v_a, v_b) < 2\epsilon$. Where the two constructions differ is in how higher-dimensional simplices are filled in.

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

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

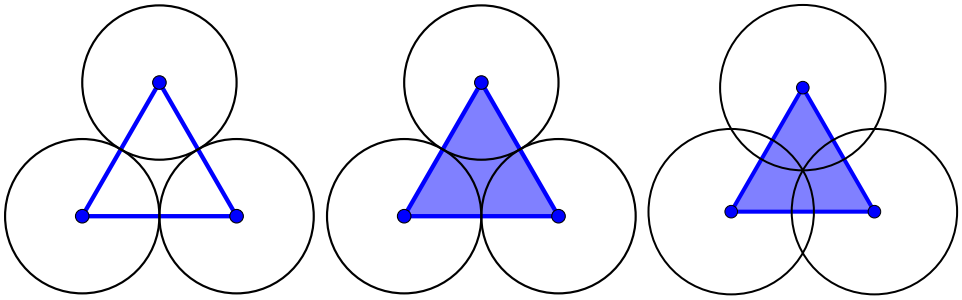
That is, the simplex $\sigma_{(s_x,\ldots,s_z)}$ is present if the intersection of balls of radius ϵ centered on vertices $(s_x, \ldots, s_z)$ is nonempty. The Vietoris-Rips complex, $VR(S, \epsilon)$, is defined as

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

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

The Čech complex is theoretically preferable because it comes with a *Nerve theorem*, which essentially states that the topology of the resulting complex will be equivalent to the topology of the union of balls used to create it. However, the Čech complex has drawbacks that prevent it from being widely applied to real data. First, computing the intersection of arbitrary balls is an expensive operation. While efficient algorithms exist in Euclidean space (the miniball algorithm [34]), it is much more difficult in arbitrary metric spaces. Furthermore, the Čech construction explicitly requires an ambient space in which the data is embedded. For data which comes in the form of a finite metric space, it may not be clear what is the ambient space.[14] In practice, the Vietoris-Rips complex is more widely applied, because it requires only the set of pairwise distances between each vertex and a scale parameter ϵ. The complex can be directly read off from the set of edges (known as the *1-skeleton*), making it extremely fast to compute.

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



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

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

2.2.2 Persistent Homology

Persistent homology is a tool developed under the umbrella of TDA that allows us to computationally study the shape of a point cloud across multiple scales simultaneously. Shape is quantified in terms of topological invariants representing homology, as discussed in the previous section. To understand why multiscale information might be of interest, consider the example in Figure 2.12. The data is sparse and noisy, but, to the eye, immediately appears to consist of two circles joined at a point along their edges. The two circles, however, are of a different radius. In the Figure, we show Vietoris-Rips complexes constructed at different scale parameters on the data. We observe that while some scale parameters are sufficient to resolve one or the other of the two circles, no single scale parameter is sufficient to simultaneously capture the two shapes. Persistent homology solves this by providing a way to track the shape information across *all* scale parameters.

Our object of study is the nested set of simplicial complexes, called a *filtration*, that is produced by tuning the scale parameter up to some threshold. At the smallest scale, $\epsilon = 0$, the complex consists only of disconnected points. As the scale parameter is increased, the

topology of the complex changes – clusters merge, holes and loops form, other holes and loops are filled – until the complex is fully connected. Each aspect of shape represents a topological invariant, and as the scale is changed, the birth and death of different invariants encoded as an interval (b_i, d_i).

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

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

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

[15]The obvious question of how to rigorously determine what makes a good interval is an open question

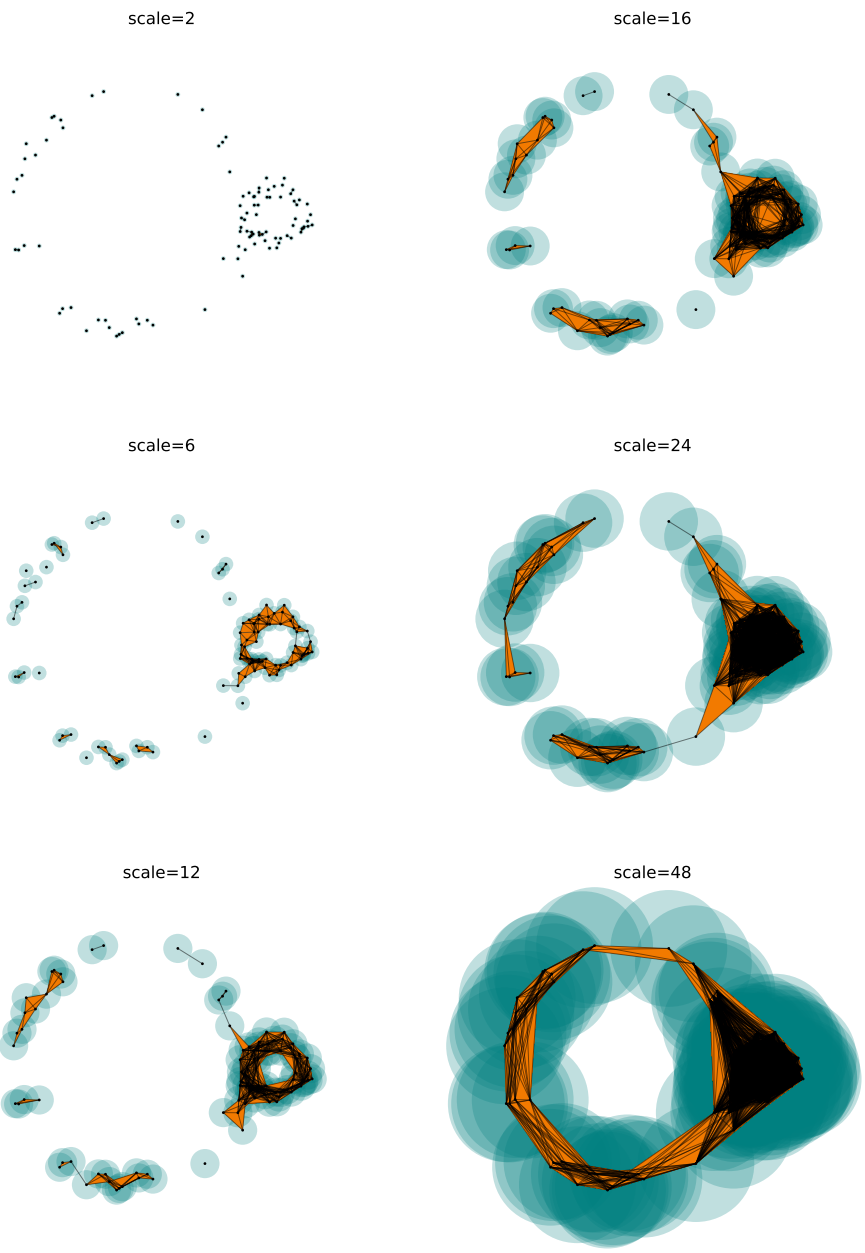


Figure 2.12: An example of constructing a filtration. The nested series of complexes form a filtration. Persistent homology will compute and track the homology at each scale. Adapted from Lesnick.

features of the data, we examine all scales simultaneously.

In fact, the persistence algorithm is more powerful than that, and can return not only

---

that is currently being addressed by a number of different groups.We discuss this further in Section 2.2.2.2.

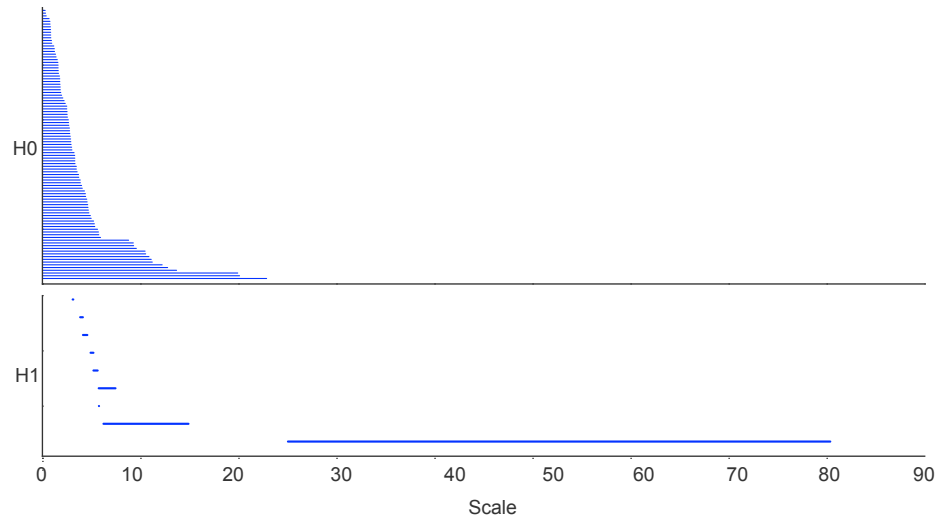


Figure 2.13: Barcode Diagram for the example in Figure 2.12. H_0 represents connectivity. H_1 represents the holes. Two holes are present at two different scales. Also some noise in the data.

the intervals associated with the invariants, but *representative cycles* of each invariant. The representative cycles correspond to the set of simplicies that surround an invariant, and can be used to determine which data points are somehow involved in a particular invariant.

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

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

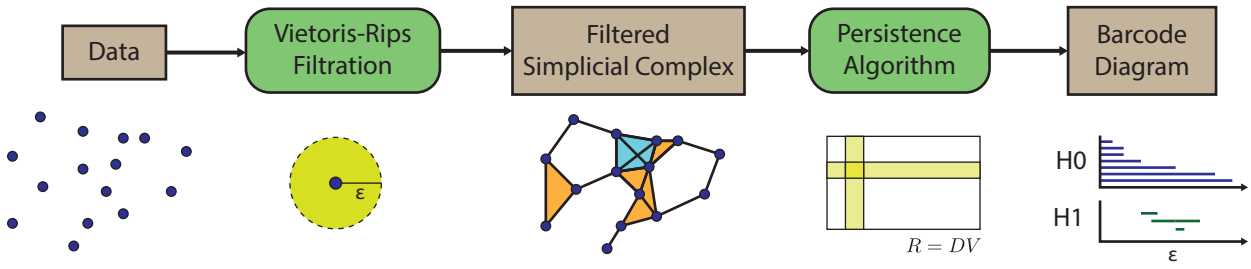


Figure 2.14: The Persistence Pipeline.

are largely beyond the scope of this thesis. Effectively, it involves manipulating the boundary matrix into a particular reduced form, from which each bar and representative cycle can be read off. Several packages for computing persistent homology have been developed, including Javaplex [61], Dionysus [51], Perseus [53], Gudhi [49], and PHAT [6]. Additionally, there is a TDA package for R which wraps functions from Dionysus and Gudhi in a user-friendly frontend [30].[16]

2.2.2.1 Stability of the Persistence Algorithm

While not strictly necessary in this thesis, the stability result is important

While not directly utilized in this thesis, the statement of stability is important to include for While not directly utilized in this thesis, the statement is important to include. The stability result establishes that small perturbations in the data will produce only small changes in the the persistence diagram. The stability result is

The result of Chazal, Cohen-Steiner, Guibas, Mémoli, and Oudot [18] 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.

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

[16]In our work we have relied on a variety of these packages. For straight-forward construction of the barcode diagram, we find the R package TDA easiest to use. If one needs to directly build and manipulate filtered simplicial complexes, Dionysus has convenient Python bindings. For large datasets, PHAT and it's parallel implementation DIPHA [4, 5] are recommended.

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 [18]. After a few definitions, we state the stability theorem.

Definition 1. 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 2. 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*.

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

Theorem 3. *The stability theorem.*

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

2.2.2.2 Statistical Persistent Homology

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

However, this leaves open the question of determining how long a bar must be to be considered significant. How short is short, and how will noisy sampling effect the observed diagram? When can a long interval be interprested as a real feature? In general, one would like to be able to develop statistics from the barcode diagram and assign measures of confidence to our estimates. How to use encode the barcode diagram in such a way that statistics can be used.

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

There are two main approaches to statistical peristent homology. The first computes functional summaries of the barcode diagram, which can then be used in downstream in a machine learning setting. In the first, functional summaries of the persistence diagram are computed. These functional summaries can be fit to known distributions and used to make inferences. In the second, probability measures on the space of persistence diagrams are directly computed. These approaches require the space of persistence diagrams to satisfy certain properties, such as being a Polish space. A Polish space has a well defined notion of mean and variance.

- 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

Fasy and coauthors have developed ways of generating confidence intervals for persistence diagrams [29]. 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 [11, 10]. 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.

Also see the work of Turner [62], Mileyko [50], and Mukherjee.

2.2.2.3 Multidimensional Persistence

First laid out in [15]. More work in [46]. 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.]

2.2.3 Mapper

Mapper sits within the large universe of dimensionality reduction algorithms for exploratory data analysis. Mapper allows for qualitative analysis of high-dimensional data through direct visualization. In this sense it belongs within the larger category of dimensionality reduction techniques such as multidimensional scaling (MDS) and their nonlinear extensions, including Isomap and t-SNE.

Mapper has the following advantages:

(1) Coordinate free (2) Invariance to deformation - robustness to noise (3) Compressed representation - ability to handle large datasets.

The *Mapper* algorithm was developed by Gurjeet Singh and Gunnar Carlsson in [59]. Further exposition can be found in [47]. Mapper was first applied to problems in RNA folding in [9] and breast cancer subtype identification in [55].

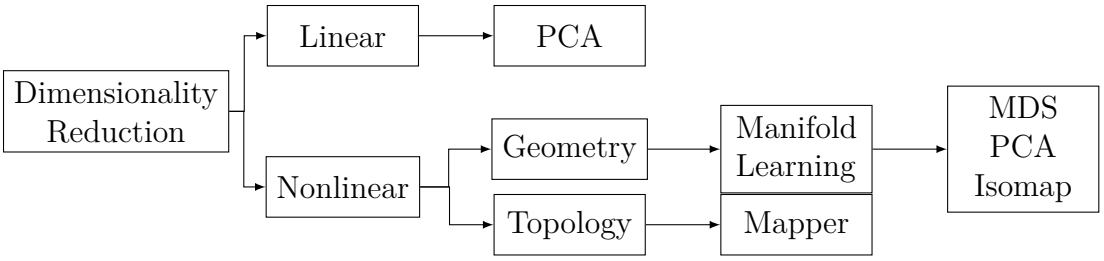


Figure 2.15: Dimensionality Reduction for EDA

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

Our use of the Mapper algorithm will be relatively minimal. We use it as a means to visualize our sequence data in such a awy In our work we use the commercial implementation of Mapper developed by Ayasdi [42]. An open-source implementation of Mapper is available in the Python Mapper package [52].

2.3 Applying TDA to Molecular Sequence Data

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

The process of evolution can be seen as an exploration of sequence space. An individual genotype can be assigned a fitness, w, which will describe the probability of reproductive success in a particular environment. Clonal evolution is the process of smoothly moving through sequence space, while reticulate evolution is the process of making discontinuous jumps through the space.

Our data generally consists of a subset of points sampled from sequence space. These points reflect a particular evolutionary history. As more data is acquired, areas of sequence space will become more densely sampled and the ability to reconstruct that evolutionary history will become more feasible. Given molecular sequence data, our program is to (1) encode the data as a finite metric space, (2) use tools from TDA to characterize the topology of the data, and (3) interpret the topology in an evolutionary context. In particular, we apply persistent homology, and read phylogenetic information contained in the dataset off the resulting barcode diagram. We make two preliminary remarks before considering a more complete example.

2.3.1 Topology of Tree-like Metrics

An important foundational point was demonstrated by G. Carlsson in [17]. Recall that tree-like data will have an additive metric space, as described in Section 2.1.4.2. In [17], it was proven that for additive metric spaces, the Vietoris-Rips filtration of the data will consist of a nested set of acyclic complexes. Consequently, the persistent homology of additive data will have nontrivial topology only in dimension zero. Furthermore, while noise in the data an introduce small deviations from additivity, the theorem puts bounds on the size of the topological features that can arise in this manner. These bounds rely on the Gromov-Hausdorff stability conditions described in Section 2.2.2.1.

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

2.3.2 The Fundamental Unit of Reticulation

In population genetics, there is a simple test for the presence of reticulate evolution in sequence data called the *four-gamete test* [40]. The test assumes only an infinite-sites model,

which states that for a sufficiently long genome, a particular residue can only ever undergo a single mutation. Put another way, there is no multiple-mutation or back mutation. The infinite-sites model has three consequences: first, one need only consider segregating sites, or nucleotide positions that have undergone a mutation. Second, because a given position can mutate only once, it is sufficient to represent sequences as binary strings, where a 0 indicates the unmutated state and 1 the mutated state. Third, for a given position we can arbitrarily assign the unmutated and mutated states. The infinite-sites model is considered a reasonably good model for long genomes.

The four-gamete test identifies reticulate evolution by looking at pairs of segregating sites. Given biallelic data, there are four possible haplotype patterns, or states, for a pair of segregating sites: 00, 10, 01, or 11.[17] The statement of the four gamete test is this: in any given dataset, the simultaneous presence of all four haplotype states in any pair of segregating sites is incompatible with strictly clonal evolution, and indicates reticulate evolution. To see this, assume state 00 as the ancestor to states 10 and 01, which arise from two independent mutations. Because of the no multipe-mutation assumption, it is not possible for either of these two states to then independently mutate into state 11. The only way for state 11 to arise is via a reticulate event that brings together the left site from state 10 and the right site from 01.[18] This process is illustrated in Figure XXA.

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

$$d = \begin{pmatrix} 0 & 1 & 1 & 2 \\ 1 & 0 & 2 & 1 \\ 1 & 2 & 0 & 1 \\ 2 & 1 & 1 & 2 \end{pmatrix} \tag{2.14}$$

The Vietoris-Rips filtration of this space is shown in Figure 2.16C. At $\epsilon = 0$ the four sequences

[17] These sites need not be adjacent.

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

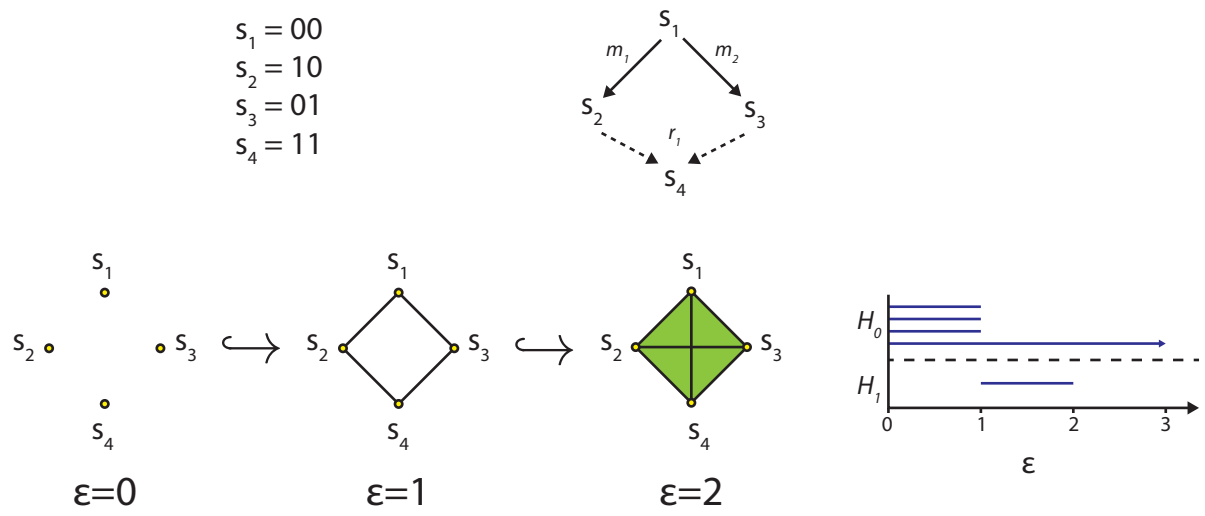


Figure 2.16: The Fundamental Unit of Reticulation. (A) A set of four sequences. (B) An evolutionary genealogy including two mutations (m_1 and m_2) and a single reticulation (r_1). (C) Vietoris-Rips Filtration. (D) Barcode Diagram.

are disconnected. At $\epsilon = 1$, four edges are drawn, forming a loop. At $\epsilon = 2$, the space is completely connected and the loop is killed. Persistent homology captures the presence of this loop as an H_1 feature in the interval $[1, 2)$ (Figure 2.16D). In this way, the reticulate event is associated with the presence of a nonzero H_1 bar.

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

2.3.3 A Complete Example

We illustrate a complete example of how TDA can capture horizontal evolution from complex populaton data in Figure 2.17. Consider the reticulate phylogeny (Figure 2.17A): 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 2.17B, 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 2.17C we demonstrate the construction of a

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

filtered simplicial complex, showing how the connectivity changes as the scale parameter ϵ is increased. Finally, in Figure 2.17D 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 of reticulation in genomic data sets.

We summarize the connection between genomic data and TDA in Table 2.1.

2.3.4 The Space of Trees, Revisited

In Section 2.1.4.4, tree space was introduced as an abstract construction to systematically represent the set of all possible binary trees. Further, we have shown that additive metrics can be represented as binary trees in tree space. Because real sequence data will very rarely

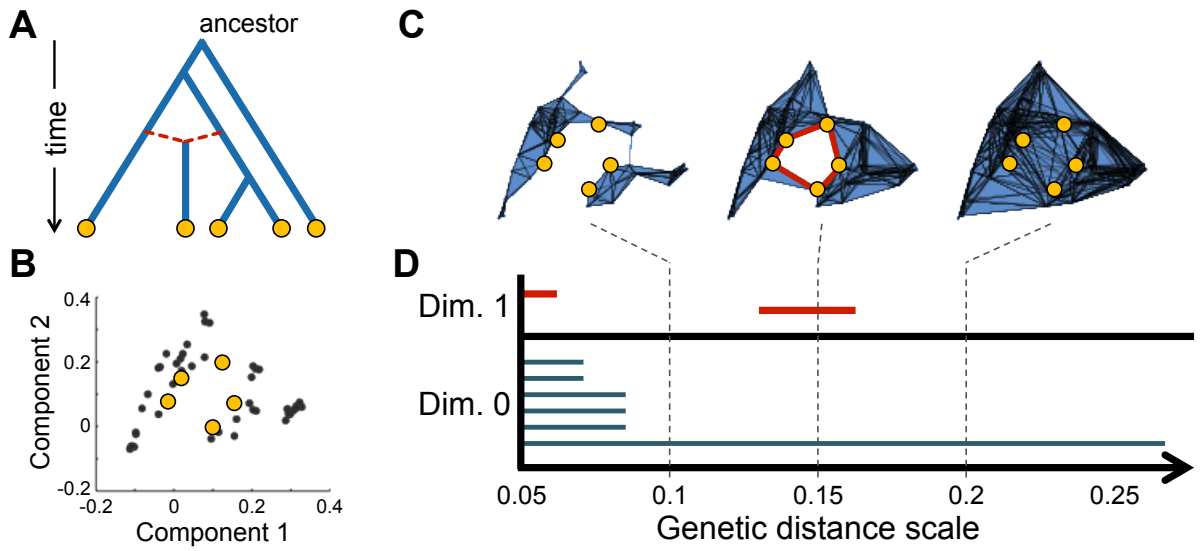


Figure 2.17: Applying persistent homology to genomic data. (A) An evolutionary genealogy including reticulation. (B) Data projected into 2-dimensions. (C) Construction of a filtered simplicial complex. (D) The resulting multiscale barcode diagram.

satisfy the additivity condition, one possible interpretation of phylogenetic reconstruction is of finding the best projection onto tree space for arbitrary data.

The complete space of finite metrics on L vertices is $(\mathbb{R}^{\geq 0})^{\binom{L}{2}}$ Tree space will be a set of $(2L-5)!!$ subspaces of dimension $\mathbb{R}^{L-3}$.

The program we propose can be understood as an extension of the tree space framework. Rather than attempt to characterize arbitrary data as a projection onto tree space, we will use methods of TDA to compute topological invariants. The intuition we develop is that as we move off of tree space, there is no additive tree that can represent the data. We would like a topological measure of deviation from addivity. Luckily, from the theorem due to Carlsson, we know we can use homology as this measure. Higher homology will vanish for additive, tree-like, metrics. The hypothesis is that the further the data lives from tree space, the larger or more these invariants will be. The deviations from additivity are telling us something about the scale and frequency of reticulate evolution in the dataset. Our updated picture looks as Figure 2.18. We have tree space, however it is now e Tree space is a subspace of

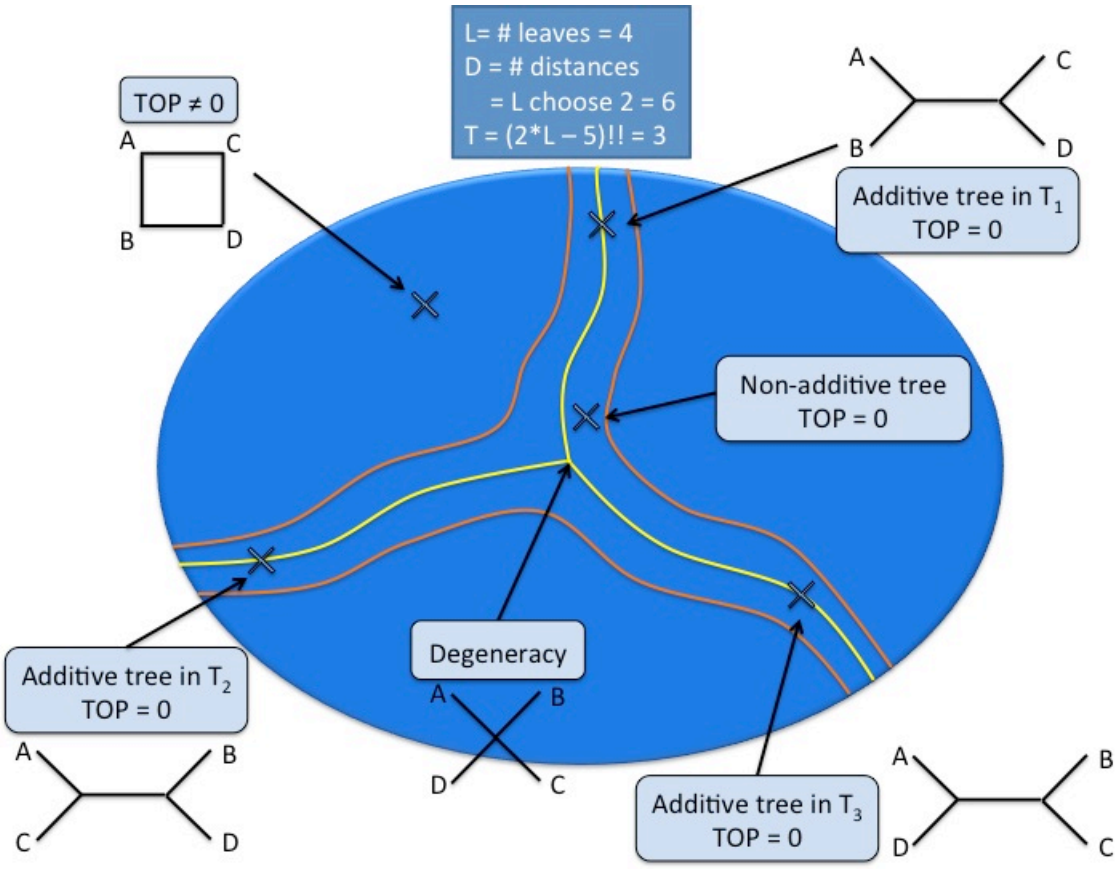


Figure 2.18: In light of reticulate evolution, our domain of interest can be expanded from the original tree space construction. We are now interested in data that lies off of tree space. We will use topology and TDA to characterize the deviations from additivity.

Note that this is an abstract construction and for data sets with an appreciable number of leaf nodes it will be difficult to directly visualize their position. Our hope is that the topological tools we develop here are sensitive enough to capture these deviations in interpretable ways.

Finally, we note that at the outset of this work, an ambitious goal was set to provide a complete characterization of the space of finite metrics in terms of their topological invariants as measured by persistent homology. While some interesting work has explored the combinatorial structure of the space of metrics on low numbers of points (see [60]), it does not appear feasible in general to provide a complete decomposition.

Bibliography

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

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

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

[4] U. Bauer, M. Kerber, and J. Reininghaus, *DIPHA (a distributed persistent homology algorithm)*, version 2.1.0, 2014.

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

[6] U. Bauer, M. Kerber, J. Reininghaus, and H. Wagner, *Phat: Persistent homology algorithms toolbox*, version 1.4.0, 2015.

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

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

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

[10] P. Bubenik, "Statistical topological data analysis using persistence landscapes," *Journal of Machine Learning Research*, vol. 16, pp. 77–102, 2015.

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

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

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

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

[15] G. Carlsson and A. Zomorodian, "The theory of multidimensional persistence," *Discrete & Computational Geometry*, vol. 42, no. 1, pp. 71–93, 2009. DOI: 10.1007/s00454-009-9176-0.

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

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

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

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

[20] F. Crick, "Central dogma of molecular biology," *Nature*, vol. 227, pp. 561–563, 1970. DOI: 10.1038/227561a0.

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

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

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

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

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

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

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

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

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

[30] B. Fasy, J. Kim, F. Lecci, C. Maria, and V. Rouvreau, *Tda: Statistical tools for topological data analysis*, version 1.4.1, 2015.

[31] J. Felsenstein, *Inferring Phylogenies*. Sunderland, Massachusetts: Sinauer Associates, 2004.

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

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

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

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

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

[37] S. J. Gould, *The Structure of Evolutionary Theory*. Harvard University Press, 2002.

[38] A. Hatcher, *Algebraic Topology*. Cambidge University Press, 2002.

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

[40] R. R. Hudson and N. L. Kaplan, "Statistical properties of the number of recombination events in the history of a sample of dna sequences," *Genetics*, vol. 111, no. 1, pp. 147–164, 1985.

[41] J. Huxley, *Evolution: The Modern Synthesis*. MIT Press, 1942.

[42] A. Inc., *Ayasdi core*, 2015. [Online]. Available: http://www.ayasdi.com.

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

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

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

[46] M. P. Lesnick, "Multidimensional interleavings and applications to topological inference," PhD thesis, Stanford University, 2012.

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

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

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

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

[51] D. Morozov, *Dionysus: A C++ library for computing persistent homology*, 2012.

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

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

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

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

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

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

[58] N. Saitou and M. Nei, "The neighbor-joining method: A new method for reconstructing phylogenetic trees," *Molecular Biology and Evolution*, vol. 4, no. 4, pp. 406–425, 1987.

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

[60] B. Sturmfels and J. Yu, "Classification of six-point metrics," *Electronic Journal of Combinatorics*, vol. 11, R44, 2004.

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

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

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

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

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

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

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

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

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

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