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 [10] (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 [4]. 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 [2]. 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.

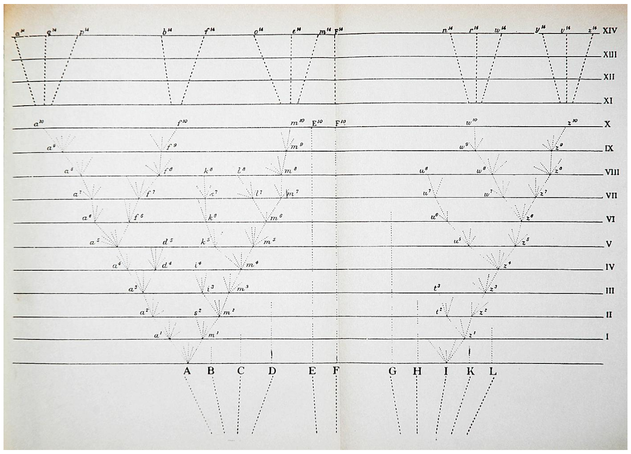
a^{14} q^{14} p^{14} b^{14} f^{14} o^{14} e^{14} m^{14} F^{14} n^{14} r^{14} w^{14} y^{14} v^{14} z^{14}
XIV
XIII
XII
XI
a^{10} f^{10} m^{10} E^{10} F^{10} w^{10} z^{10}
X
IX
VIII
VII
VI
V
IV
III
II
I
A B C D E F G H I K L

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 [10].

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 [12, 27].

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 [30]. 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 [20] and [18] for historical detail.

[3] In humans, only 1.5% of the genome is protein-coding, the rest largely non-functional. [23]. 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 [34, 35]. 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 [16]. 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 [32]. 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).

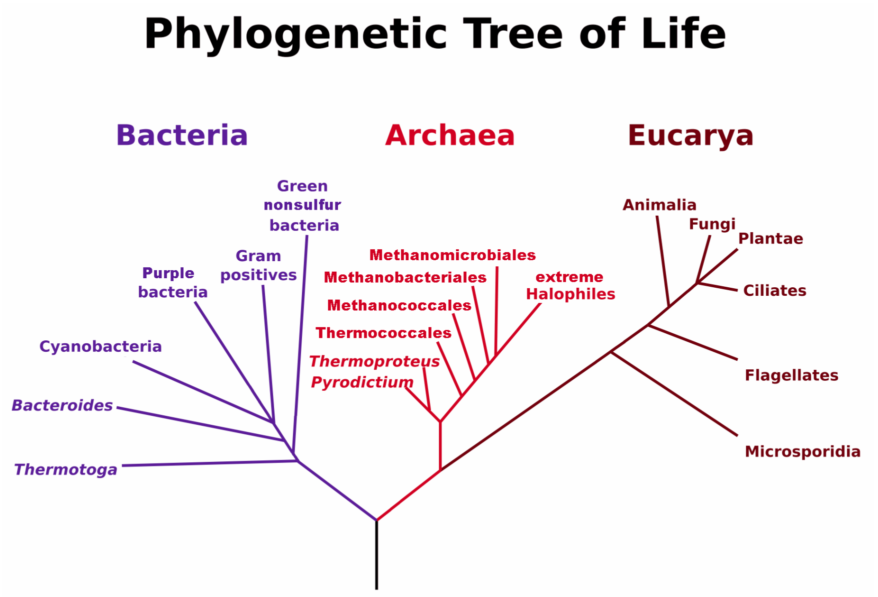

Figure 1.2: Carl Woese's three kingdom tree of life. Using 16S subunit ribosomal RNA, Carl Woese identified archaea as a distinct phylogenetic kingdom. Previously, based on morphological similarity (specifically, unicellular and lacking a nucleus), archaea had been grouped with bacteria. This result was an early success for molecular phylogenetics and the use of conserved gene segments for molecular classification. Figure adapted from [33].

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 [31, 17]. 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 [9]). 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 [7], 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 [26]. Incompatibilities in the tree paradigm now appear as the rule, not the exception, which has led to calls for new representations of evolutionary relationships [12, 13]. Many have argued that, in light of new genomic evidence, the very notion of a universal tree of life must be discarded [21, 22].

[5] There exists a cottage industry of methods for aggregating conflicting *gene trees* into a consensus *species tree*, see [24].

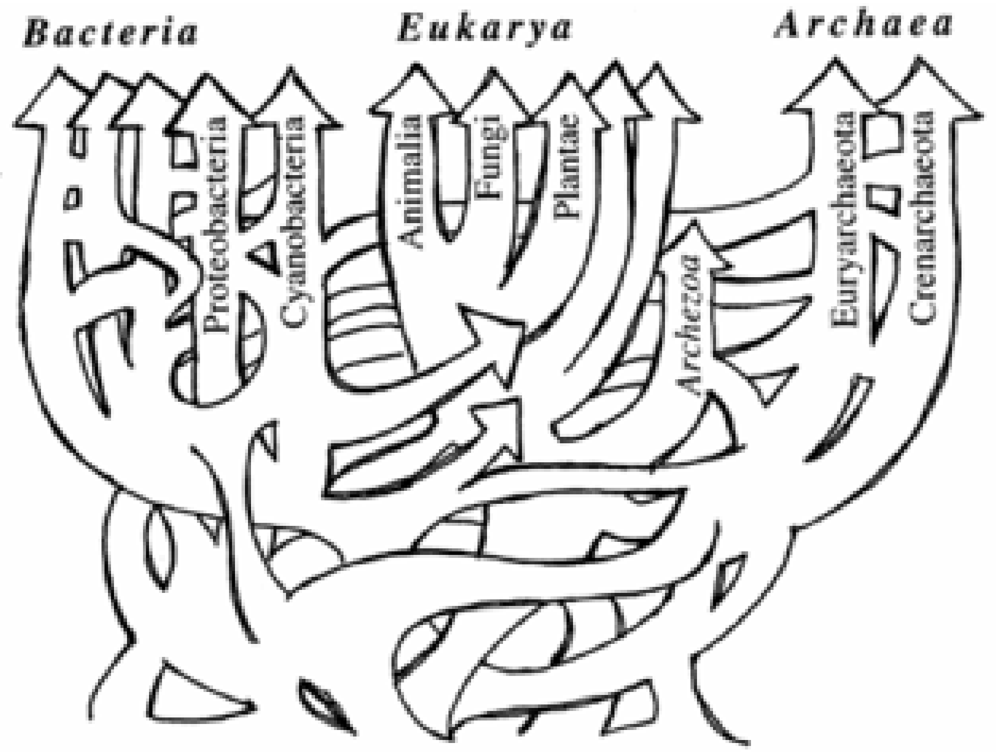

Figure 1.3: W Ford Doolittle's representation of the universal tree of life with reticulate evolution. While the three domains of life are still recognizable, patterns of divergence no longer follow a strictly treelike model. (From *Science*, vol. 284, issue 5423, page 2127. Reprinted with permission 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 [5]. In influenza, segmental gene reassortments lead to antigenic novelty and the emergenence of epidemics [25]. In several pathogenic bacteria, including *E. coli* and *S. aureus*, horizontal gene transfer has been responsibile for the spread of antibiotic resistance genes [1, 11].

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

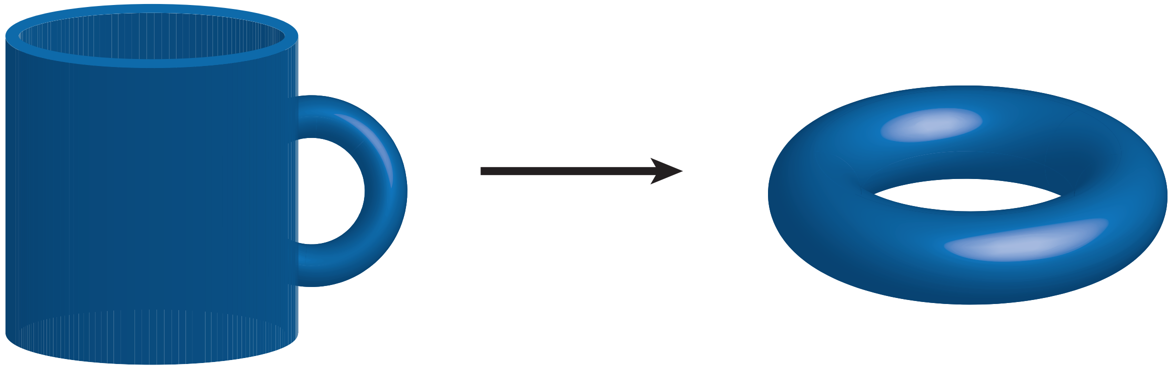

Figure 1.4: The paradigmatic example of topological equivalence. The coffee mug can be continuously deformed into the donut and are therefore topologically equivalent. Both exhibit the topology of a solid torus ($D^2 \times S^1$).

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

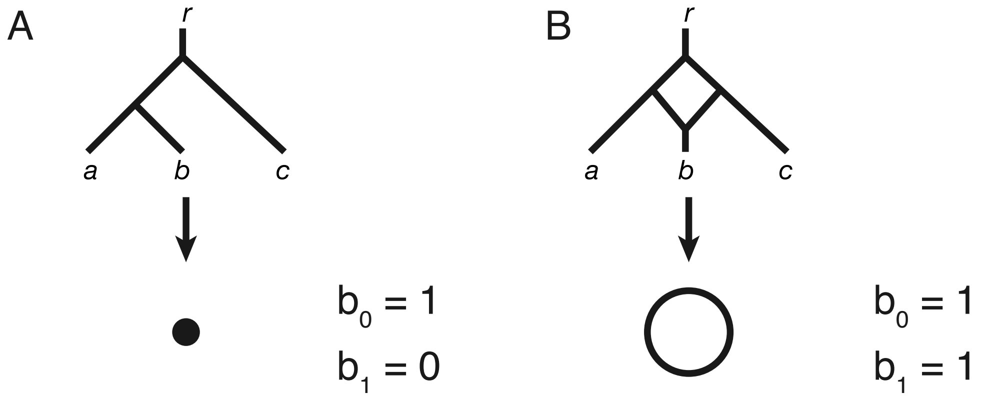

Figure 1.5: (A) A simple treelike phylogeny is contractible to a point. (B) A reticulate phylogeny that is equivalent to a circle and not contractible without a cut. The two spaces are not topologically equivalent and can be distinguished by their Betti numbers.

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 [14]. 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 [15]

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 open problems in evolutionary biology and genomics. Here we supply sufficient background to motivate our approach Exposition for specific applications can be found in their respective individual chapters.

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 phylogenetic methods by which it is analyzed.

2.1.1 Genes and Genomes

The information required to express an organism's biological form and function is contained in the genome, the complete sequence of nucleotides (DNA) contained in every cell[1] Embedded in this sequence are subsequences representing genes, which code for the protein products

[1] While often represented as a linear string of characters, this representation can be misleading, as many organisms, primarily viruses and bacteria, have circular genomes.

Embedded in the genome is a complex regulatory pattern of transcription factors controlling the expression of particular genes controlling cell differentiation and development. that controls the differential expression of particular Embedded in this sequence is a complex regulatory pattern of transcription factors controlling the expression of particular genes.

Include a figure of a gene / genome? Following the central dogma of biology, DNA is transcribed into RNA, RNA is translated into amino acids, and amino acids are folded into proteins [8]. 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 the observed diversity. The field concerned with exploring these relationships is *comparative genomics*.

2.1.2 Evolutionary Processes

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

2.1.2.1 Clonal Evolution

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

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

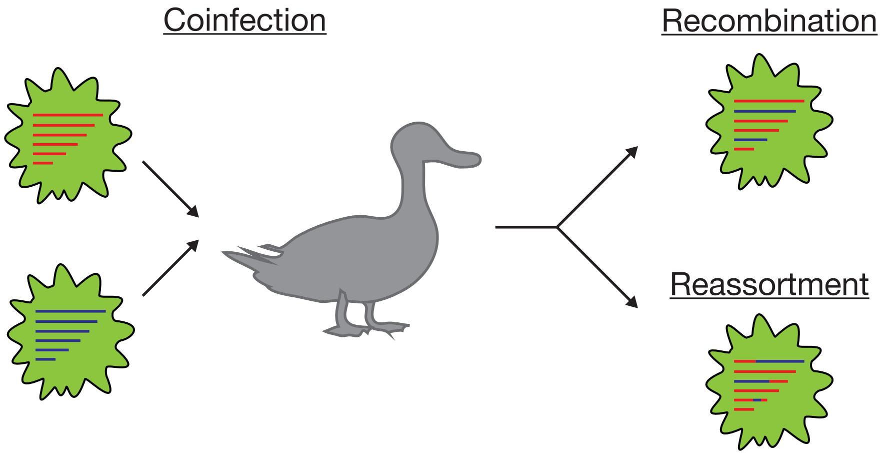

Figure 2.1: The two modes of viral reticulation. Coinfection of the same host cell can lead to either reassortment, where whole viral segments are exchanged, or recombination, in which breakpoints can occur within segments. The former process is common in influenza, the latter in HIV. The end result, however, is

2.1.2.2 Reticulate Evolution

Reticulate evolution, known also as 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 of evolutionary relationships difficult. Different reticulate processes occur in different types of organisms (summary in Table 2.1.2.2).

In viruses, reticulation can occur when two virus particles coinfect the same host cell. In the process of replicating, the viruses can exchange genetic material in one of two ways: reassortment and recombination (the two processes are contrasted in Figure 2.1). Reassortment occurs in viruses whose genomes are structured into distinct segments, such as influenza. Segments can be thought of as chromosomes. Viral replication involves the packaging of each segment of the virus, and in the presence of coinfection segments from different virus particles can exchange. Recombination occurs in nonsegmented virusesm, such as HIV, and involves intrasegmental crossover. Intrasegmental crossover is a type of *homologous recombination*. Reassortment, on the other hand, is a type of *nonhomologous recombination*.

In bacteria, reticulation can often involve the direct exchange of genes. Reticulation is often called horizontal gene transfer or lateral gene transfer. Horizontal exchange occurs when a donor bacteria transmits foreign DNA into a genetically distinct bacteria strain. Three mechanisms of horizontal transfer are identified, depending on the route by which foreign DNA is acquired [26]. Foreign DNA can be acquired via uptake from an external environment (transformation), via viral-mediated processes (transduction), or via direct cell-to-cell contact between bacterial strains (conjugation).

In eukaryotes, several reticulation processes

[2]

In eukaryotes: hybridization and meiotic recombination. Patterns of meiotic recombination can be

Organism	Process	Description
Virus	Reassortment	INSERT FIGURE
	Recombination	INSERT FIGURE
Bacteria	Transformation	xx
	Transduction	xx
	Conjugation	xx
Eukaryotes	Meiotic Recombination	Description
	Hybridization	Description

The presence of horizontal gene transfer in a set of organisms can be most clearly identified by comparing the phylogenetic trees built from different genes. If horizontal gene transfer has occurred, the set of *gene trees* will reflect different evolutionary relationships and not be consistent with a single tree topology. A subfield of comparative genomics is concerned with building *species trees* from sets of gene trees, however in the case where

[2] One might wonder about sexual reproduction – a descendent can be seen as a hybridization of both mother and father. Indeed, sexual reproduction can itself be considered a form of reticulate evolution. In the case of *H. sapiens*, a diploid species with two copies of each chromosome, XXX. However, from an evolutionary perspective, the process can be decoupled into two independent clonal processes. When discussing the human species, researchers will commonly refer to the most recent common ancestor on the patrilineal as Y-chromosomal Adam and the most recent common ancestor on the matrilineal as mitochondrial Eve.

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

2.1.3.1 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. [Figure comparing Wright-Fisher and Coalescent.]

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

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

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

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

The coalescent with recombination is an extension of this model that allows different genetic loci to have different genealogies. Looking backward in time, recombination is modeled as a splitting event, occurring at a rate determined by population-scaled recombination rate ρ, such that an individual has a different ancestor at different loci. Evolutionary histories are no longer represented by a tree, but rather by an *ancestral recombination graph.* Recombination is the component of the model generating nontrivial topology by introducing deviations from a contractibile tree structure, and is the component which we would like to quantify. Coalescent simulations were performed using `ms` [19].

2.1.3.3 Metrics on Seqeuces

What metrics can we put on aligned sequences?

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

The simplest model, and the one most commonly adopted in this thesis, is the Hamming metric, which simply counts the differences between two aligned sequences.

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

Jukes-Cantor metric is defined as

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

2.1.4 Phylogenetic Methods

A phylogenetic tree is XXX.

Molecular phylogenetics refers to a large collection of methods for inferring evolutionary relationships among a set of species from molecular sequence data. In practice, this In practice: tree-building.

Starting with a set of sequences that share some similarity, an alignment is performed. The alignment allows columns of the sequence to be directly compared. From an alignment, one can then directly use parsimony or likelihood approaches. Alternatively, one can compute a matrix of pairwise distances and then construct a tree that best approximates these distances. Most relevant to this thesis are the distance-based approaches (because they can be viewed as finite metric spaces amenable to topological analysis). Only in the case of perfectly additive data will a tree be able to exactly fit the matrix. (Define additivity – four point condition.) Identifying a pairwise distance matrix with a finite metric space representation is a crucial step that allows most of the machinery described later to be applied. Rooted vs unrooted.

2.1.4.1 Distance Matrix Methods

Introduced by Cavalli-Sforza and Edwards in 1967 [6] and Fitch and Margoliash in 1967 [16]. Fitch-Margoliash method is a weighted least squares tree-fitting method (larger distances are weighted less, due to higher random error). Compute a matrix of pairwise distances and then find the tree that best approximates those distances. Neighbor joining is now the most commonly used distance-matrix approach because it can perfectly reconstruct an additive tree. Neighbor joining was introduced by Saitou and Nei in 1987 [28].

Limitation: Distance methods do not make use of all of the information in the sequence.

2.1.4.2 Phylogenetic Networks

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

2.1.4.3 Number of Tree Topologies

The number of unrooted bifurcating tree topologies with L leaves is $(2L-5)!!$.[3] 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. Fitch quote: *more than 20 species, more than Avogadro's number of topologies.*

2.1.4.4 Space of Phylogenetic Trees

A phylogenetic tree is characterized by its number of leaves, the particular tree topology, and the length of each branch. Tree space refers to an abstract construction that represents each possible tree as a point in a geometric space. Abstract studies of tree space were initiated by Billera, Holmes, and Vogtmann (BHV) in [3].

In the BHV model, each point represents an unrooted binary tree with L leaves and strictly positive branch lengths.

Number of interior edges $r = L-3$, a particular additive tree can be plotted as a point in the positive open orthant $(0,\infty)^r$.

A single orthant corresponds to a single tree topology.

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

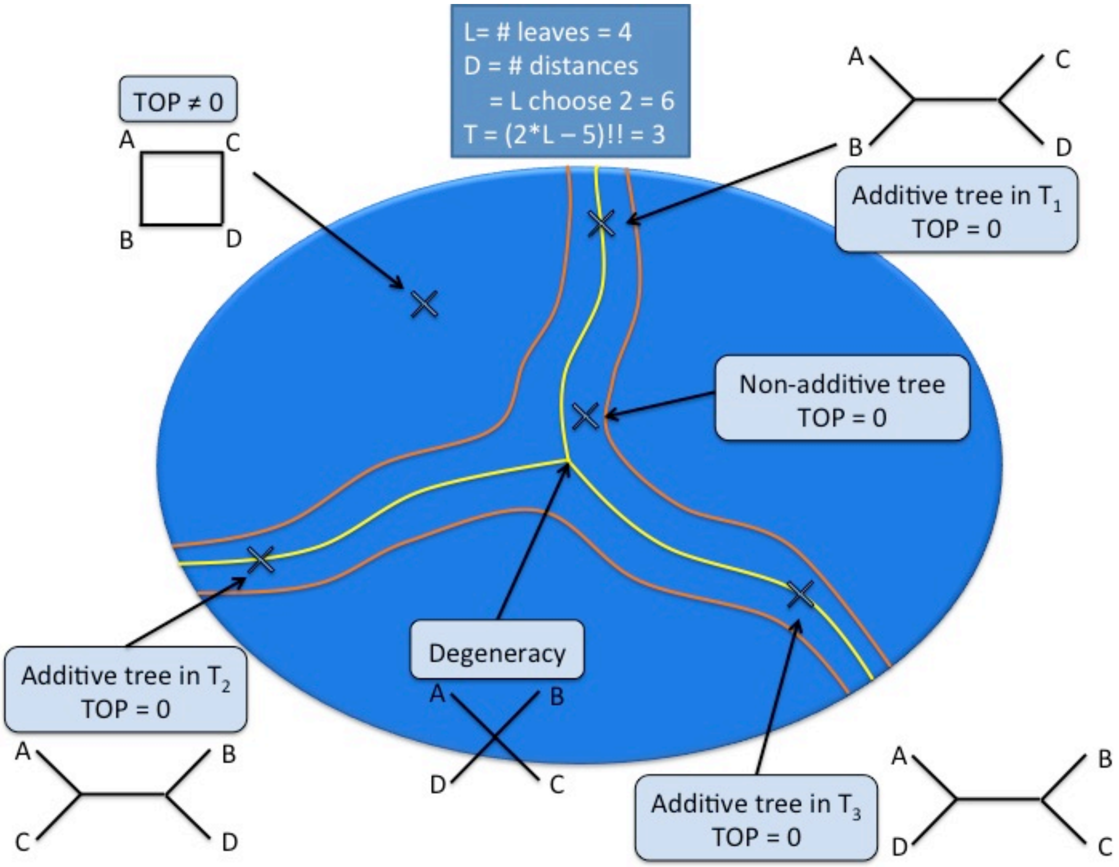

Figure 2.2: Tree Space

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.

[2] M. L. Arnold, *Evolution through Genetic Exchange*. Oxford University Press, 2007.

[3] 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.

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

[5] D. Burke, “Recombination in hiv: an important viral evolutionary strategy,” *Emerging Infectious Diseases*, vol. 3, no. 3, pp. 253–259, Sep. 1997.

[6] 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.

[7] 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.

[8] F. Crick, “Central dogma of molecular biology,” *Nature*, vol. 227, pp. 561–563, 1970.

[9] T. Dagan and W. Martin, “The tree of one percent,” *Genome Biology*, vol. 7, no. 10, p. 118, 2006.

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

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

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

[17] N. Goldenfeld and C. Woese, "Biology's next revolution," *Nature*, vol. 445, no. 7126, pp. 369–369, Jan. 2007.

[18] S. J. Gould, "The structure of evolutionary theory," 2002.

[19] R. R. Hudson, "Generating samples under a wright–fisher neutral model of genetic variation," *Bioinformatics*, vol. 18, no. 2, 2002.

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

[21] E. V. Koonin, "Darwinian evolution in the light of genomics," *Nucleic Acids Research*, vol. 37, no. 4, pp. 1011–1034, Dec. 2008.

[22] 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.

[23] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J. P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, N. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. M. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J. C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R. H. Waterston, R. K. Wilson, L. W. Hillier, J. D. McPherson, M. A. Marra, E. R. Mardis, L. A. Fulton, A. T. Chinwalla, K. H. Pepin, W. R. Gish, S. L. Chissoe, M. C. Wendl, K. D. Delehaunty, T. L. Miner, A. Delehaunty, J. B. Kramer, L. L. Cook, R. S. Fulton, D. L. Johnson, P. J. Minx, S. W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning,

T. Slezak, N. Doggett, J.-F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, R. A. Gibbs, D. M. Muzny, S. E. Scherer, J. B. Bouck, E. J. Sodergren, K. C. Worley, C. M. Rives, J. H. Gorrell, M. L. Metzker, S. L. Naylor, R. S. Kucherlapati, D. L. Nelson, G. M. Weinstock, Y. Sakaki, A. Fujiyama, M. Hattori, T. Yada, A. Toyoda, T. Itoh, C. Kawagoe, H. Watanabe, Y. Totoki, T. Taylor, J. Weissenbach, R. Heilig, W. Saurin, F. Artiguenave, P. Brottier, T. Bruls, E. Pelletier, C. Robert, P. Wincker, A. Rosenthal, M. Platzer, G. Nyakatura, S. Taudien, A. Rump, D. R. Smith, L. Doucette-Stamm, M. Rubenfield, K. Weinstock, H. M. Lee, J. Dubois, H. Yang, J. Yu, J. Wang, G. Huang, J. Gu, L. Hood, L. Rowen, A. Madan, S. Qin, R. W. Davis, N. A. Federspiel, A. P. Abola, M. J. Proctor, B. A. Roe, F. Chen, H. Pan, J. Ramser, H. Lehrach, R. Reinhardt, W. R. McCombie, M. de la Bastide, N. Dedhia, H. Blöcker, K. Hornischer, G. Nordsiek, R. Agarwala, L. Aravind, J. A. Bailey, A. Bateman, S. Batzoglou, E. Birney, P. Bork, D. G. Brown, C. B. Burge, L. Cerutti, H.-C. Chen, D. Church, M. Clamp, R. R. Copley, T. Doerks, S. R. Eddy, E. E. Eichler, T. S. Furey, J. Galagan, J. G. R. Gilbert, C. Harmon, Y. Hayashizaki, D. Haussler, H. Hermjakob, K. Hokamp, W. Jang, L. S. Johnson, T. A. Jones, S. Kasif, A. Kaspryzk, S. Kennedy, W. J. Kent, P. Kitts, E. V. Koonin, I. Korf, D. Kulp, D. Lancet, T. M. Lowe, A. McLysaght, T. Mikkelsen, J. V. Moran, N. Mulder, V. J. Pollara, C. P. Ponting, G. Schuler, J. Schultz, G. Slater, A. F. A. Smit, E. Stupka, J. Szustakowki, D. Thierry-Mieg, J. Thierry-Mieg, L. Wagner, J. Wallis, R. Wheeler, A. Williams, Y. I. Wolf, K. H. Wolfe, S.-P. Yang, R.-F. Yeh, F. Collins, M. S. Guyer, J. Peterson, A. Felsenfeld, K. A. Wetterstrand, R. M. Myers, J. Schmutz, M. Dickson, J. Grimwood, D. R. Cox, M. V. Olson, and Kaul..., "Initial sequencing and analysis of the human genome," *Nature*, vol. 409, no. 6822, pp. 860–921, Feb. 2001.

[24] W. P. Maddison, "Gene trees in species trees," *Systematic Biology*, vol. 46, no. 3, pp. 523–536, Sep. 1997.

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

[26] 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.

[27] 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.

[28] 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.

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

[30] J. D. Watson and F. H. Crick, "Molecular structure of nucleic acids," *Nature*, vol. 171, no. 4356, pp. 737–738, 1953.

[31] C. R. Woese, “A new biology for a new century,” *Microbiology and Molecular Biology Reviews*, vol. 68, no. 2, pp. 173–186, Jun. 2004.

[32] 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.

[33] 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.

[34] 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.

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