Chapter 1

Introduction

1.1 Genealogical and Phylogenetic and Networks

> "All DNA is recombinant DNA ... [The] natural process of recombination and mutation have acted throughout evolution ... Genetic exchange works constantly to blend and rearrange chromosomes, most obviously during meiosis...[199]"

> "Molecular phylogeneticists will have failed to find the 'true tree', not because their methods are inadequate or because they have chosen the wrong genes, but because the history of life cannot propertly be represented as a tree. [42]"

Genealogical and Phylogenetic **Networks** are graph-theoretic models of evolution that go beyond Phylogenetic **Trees**, the traditional representation of evolutionary history. Genealogical and Phylogenetic Networks incorporate non-tree-like biological events such as *meiotic recombination* that occurs in *populations* of individuals inside a *single* species, or that incorpore general *reticulation* events that occur between different species, caused for example by *lateral gene transfer or hybrid speciation.* The central algorithmic problems are to reconstruct a plausible history of mutations and non-tree-like events that generate a given set of extant, observed genomic *sequences*, and to determine the *minimum* number of biological events needed to derive the sequences.

This book primarily concerns combinatorial and algorithmic issues involved in reconstructing the evolutionary history of extant sequences observed in populations, where the sequences are generated by mutations and recombinations. However, many of the combinatorial and algorithmic results apply equally well at the phylogenetic level, i.e., to reticulate evolution of species, and we will point these out when they occur. The book is aimed broadly at computer scientists, mathematicians, and biologists. We will explain the various biological phenomena; the mathematical, population genetic and phylogenetic models that capture the essential elements of these phenomena; the resulting combinatorial and algorithmic problems that derive from those models and from biological questions that are formulated in terms of these models; the theoretical results (both combinatorial and algorithmic) that have been obtained; related software that has been developed; and the results of empirical testing of that software on simulated and real biological data. In addition, we will explain some needed combinatorial and algorithmic background for those readers who might not be familiar with particular existing results or techniques.

1.1.1 Recombination and Genealogical Networks

Nature and history, through mutation, recombination, gene conversion, genome rearrangements, lateral gene transfer, admixture of populations, selection, random drift, etc. have "conducted" a huge number of experiments in which DNA has been mixed in different ways to create a vast variety of *mosaic* or *chimeric* genome sequences in current populations. These extant mosaic sequences can be queried in a wide variety of ways to address a large number of fundamental or applied biological questions and controversies.

The key biological event (along with point mutation) that creates mosaic genomes in *populations* (individuals in a single species) over relatively short periods of time, is **meiotic recombination** which, in every meiosis, takes the two "copies" of a chromosome in an individual and produces a third *recombinant* chromosome consisting of alternating segments (usually a small number) of the two chromosomes (see Figure 1.1). Any child of that individual then inherits such a recombinant chromosome. Similarly, recombination between the two "copies" of a chromosome in the other parent creates a different recombinant chromosome which is then passed down to the child. Considering recombination in all the prior generations, it follows that the genome that any individual inherits is a mosaic mixture and reflection of the DNA of *all* of an individual's ancestors. In this way, meiotic recombination is one of the principal evolutionary forces responsible for shaping genetic variation within a single species. It allows the rapid creation of hybrid chromosomes even without mutations at individual sites. This ability to rapidly create hybrid chromosomes is believed to be an important adaptive property, and hence through natural selection, recombination (and sexual reproduction) is a feature of all major Eukaryotic species.

Pedigrees We are interested in reconstructing plausible histories of mutations and recombinations that might have derived segments of chromosomal sequences observed in current populations. Such histories are not in the form of trees, but rather in the form of networks. To begin to explain the need for networks, we consider first the related issue of family *pedigrees*.

If we trace the ancestry of a person backwards in time, their two parental

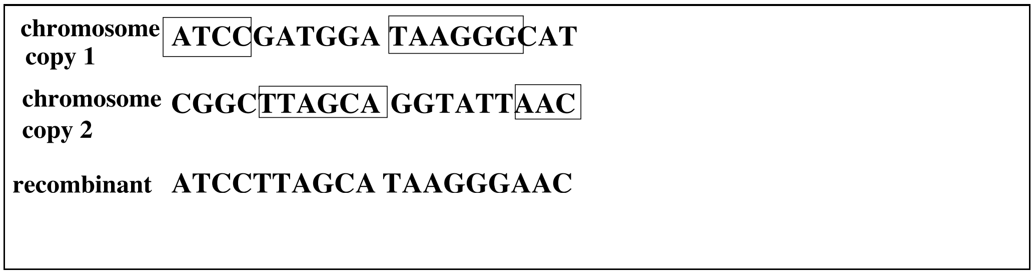

Figure 1.1: Recombination of two sequences creating a third, recombinant sequence. The recombinant chromosome is created from the boxed segments of the two chromosome copies. This example is wildly out of a scale, as the four segments are very short. In humans, the number of segments in a single recombinant chromosome is generally under ten, and so the true segments are much longer than in this example.

lines will expand into multiple lines (as parents expand to grandparents and great-grandparents, etc.) but some lines will eventually "coalesce", meaning that two distinct ancestors will have one or two common parents. See Figure 1.2. If the trace is far enough back in time, all of the ancestral lines of the individual, or a set of individuals, will coalesce to two common ancestors (one male and one female - the Adam and Eve of the sampled idividuals). It follows that the genealogical history, or *pedigree*, of a set of individuals will contain *cycles* and therefore *cannot* be represented in a *tree*; instead, the representation requires a **network**.

When we discuss recombination, it will be helpful to remember that the cycle in the pedigree of William and Harry is due to two things: first, that their father, Charles, had *two* parents, Elizabeth and Philip, and second, that the ancestral lines of Elizabeth and Philip eventually *coalece*.

Tracing the history of DNA segments with recombination We considered family pedigrees in order to introduce the notion of coalescence and ancestry cycles. But our main interest is in the history of DNA sequences rather than in families. So, we now shift attention back to chromosomes and recombination and to the cycles that arise in the generation of DNA sequences. Each individual contains two "copies" of each of several chromosomes, but we focus just on a *single* copy of a single chromosome (or chromosome segment) in a set of individuals. This is called the *haploid* case. Later, in Chapter **??**, we will consider some problems related to the *diploid* case, i.e., where we consider the joint history of both copies of a single chromosome segment in a set of individuals.

We examine the transmition history of the single chromosome segment backwards in time, and we first assume that there is *no* recombination. In that case, each individual receives the particular chromosome segment from one only par-

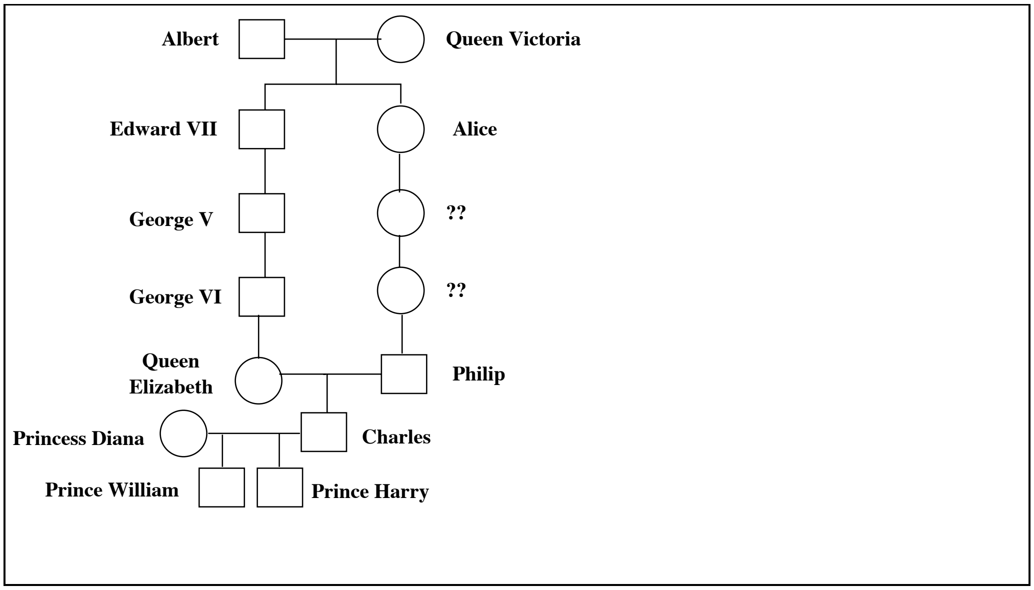

Figure 1.2: A partial pedigree of Princes William and Harry. (Note that the three females drawn opposite of Edward and the two Georges are not their mates.) William's and Harry's two parents, Princes Diana and Prince Charles, each had two parents, who each had two parents, etc. Following six generations back in time, William and Harry would have $2^6 = 32$ ancestors in that generation, if they were all distinct. (With the exception of Diana, we only show the ancestors of William and Harry who are descendants of Albert and Victoria.) But in fact, William and Harry do not have 32 ancestors in the sixth generation back from them. In the sixth generation back from William and Harry, there are two coalescent events: Edward and Alice coalesce at a common father, Albert, and they coalesce at a common mother, Victoria. That is, Edward and Alice are full siblings with the same parents. These coalescences form a *cycle* in the pedigree. Cycles of this type in the Royal English family are responsible for the high prevelance of the recessive disease Hemophilia; Victoria and Alice were carriers, but none of their shown descendants were. Note that the pedigree would also have had a cycle if Edward and Alice had only been half-siblings, i.e., if they had only shared a single parent.

ent, either their father or their mother. It follows that the transmission history of a single individual's chromosome segment forms a single *path* through the individual's pedigree. For example, Prince William might have received a copy of a chromosome segment from his father, Charles, who received the segment from Elizabeth, who received it from Albert through a path consisting of Edward, George V and George VI.

Now consider tracing, backwards in time, the transmition history of a single chromosome segment from each of *two* individuals. The transmition history of each of the two segments forms a path in the individual's pedigree, but just as the two individuals ultimately share a common ancestor in their pedigrees, ultimately the two transmission histories will coalesce at some common ancestor of the two individuals. Note that the point where the two transmission histories coalesce will be at or above the point where the the two individuals share a common ancestor in their pedigrees. More generally, the backwards transmission history of a *set* of chromosomes (one copy from each distinct individual in the set) will coalesce at one ancestor who is common to all of the individuals in the set.

Now we add in the fact of recombination, but still only consider the history of a single chromosome segment in a set of individuals. We also assume that no mutations occur in this history. However, even though we are considering the history of only a single chromosome segment (rather than histories of the two segments that each individual has), because of recombination, we have to pay attention to the fact that each individual actually contains two copies of each segment. Suppose, for example, that the segment is transmitted to the individual from the individual's father. Because of recombination, that segment might not have actually been transmitted to the father from one of his parents; rather, the segment might have been created by the father as a result of recombination (during meiosis) of the two copies of the segment that the father received from his parents.

To make the discussion concrete, consider a single copy of a chromosome segment that Prince Harry received, and assume that it was transmitted to Harry from his father, Charles. Charles might has received the segment in whole from either Elizabeth or Philip, but need not have. Instead, the segment that Charles transmits to Harry might have been created by Charles by recombining the segment Charles received from Elizabeth with the segment Charles received from Philip. We will suppose that this is the case. Further, we will assume that the segment Philip transmits to Charles is identical to the segment that Alice has, i.e., is transmitted from Alice to Charles, through two intermediate generations, without change. Similarly, we assume that the segment Charles recevied from Elizabeth was transmitted from Edward, through two generations, without change. Now to create a cycle in the historhy of Harry's chromosome segment, suppose that both Alice and Edward received an identical copy of that segment from Victoria. Then the history of Harry's segment contains a cycle which starts with Victoria and ends with Harry[1].

Because the transmition history, backwards in time, of a set of chromosome segments (one from each individual in the set) contains recombination events, and because the transmittion history ultimately coaleces at a single ancestor, the representation of that history will have cycles, and again cannot be represented by a tree; rather, its representation requires a network. This is analogous to the case of pedigrees, but different in biological detail: Prince Harry's pedigree contains a cycle because Charles has two parents whose lines eventually coalesced at a common ancestor (actually at two common ancestors, Albert and Victoria); the hypotheical history of one copy of one of Harry's chromosome segments has a cycle because of a recombination during meiosis that Charles experienced, and because the segment has a common origin with Victoria[2].

The network that represents the history of chromosome segments (one from each individual in the set) is often called a "phylogenetic network" in the computer science literature, although the term "genealogical network" is more appropriate; the network is often called an "ancestral recombination graph (ARG)" or "coalescent with recombination" in the population genetics literature. We will give a more complete discussion of network terminology in Chapter 3, where we will develop more formal definitions of a *genealogical network* and of an *ancestral recombination graph.*

1.1.2 The Central Thesis

The true genealogical network that explicitly reveals the origin and derivation of the sequences in a current population, showing the locations of all the mutations and recombinations (both in the genome and in time), tremendously facilitates the use of genome data to address biological questions of interest, particularly those involving recombination[3]. Unfortunately, we cannot directly examine the past so we cannot know (for sure) the true genealogy of the extant

[1] Note that even though Edward and Alice have the same parents, the cycle need not have started at that generation. This would happen, for example, if Alice recevied a copy of the segment from Albert and Edward received it from Albert. In that case, any cycle would start at a generation prior to Albert and Victoria.

[2] An astute reader will observe that the copies of the segments received by Elizabeth and Philip will be identical, and hence their recombination will create another identical copy. That would lead to the question of why we care about the history. In fact, point mutations also occur in the history, so that the copies that Elizabeth and Philip receive will not be indentical, and the segment that Harry receives is likely different from any other segment in the history.

[3] And recombination is of great interest throughout biology. In addition to, or through, its role in producing genetic variation, recombination is central in a large number of diverse biological phenomena (including some serious diseases), many of which may seem at first to be unrelated to recombination. A Google Scholar search (properly restricted) shows over one million articles concerning biological recombination.

sequences. However, a robust literature on *algorithms* that construct plausible genealogical networks, or deduce well-defined aspects of a genealogy, has developed, particularly in the last several years. Related questions about "hybridization networks", which are similar to genealogical networks but do not necessarily involve explicit sequences, have also been addressed. This algorithmic research has been encouraged by a growing appreciation by biologists that many evolutionary and population genetic phenomena must be represented by networks rather than by trees.

Even though we can never know for sure that an algorithm has deduced the correct genealogical network (or features of it), we will detail in this book that applications of these algorithms have correctly answered certain biological questions, suggesting that important parts of true genealogies are captured in or reflected by the computations. These applications go to the heart of the **Central Thesis** of this book, that

> *Explicit* genealogical networks representing a derivation of extant sequences in a population can be effectively computed, and even if those networks do not perfectly capture the true history they can reveal parts of the history and give significant insight into important biological phenomena.

Problems of constructing genealogical networks from sequence data, or deducing features of such networks, are significantly more complex than for the analogous problems in *trees*, and the field of network reconstruction is much less developed than the field of tree reconstruction. But as more population genomic data accumulates, problems defined on networks will become increasingly important, and with those problems the importance of efficient, scalable algorithms will increase.

This book discusses algorithmic and mathematical results, most obtained in the last decade, concerning combinatorial structure of genealogical networks (sometimes extending to other phylogenetic networks). The algorithms exploit the structure, and are used to deduce information (sometimes only partial) about the networks, or are used to explicitly construct networks that generate observed sequences through the biological events of mutation and recombination (and sometimes other events). The networks serve as hypotheses for the true genealogical history of the extant sequences, and help to address fundamental biological questions, or are used in practical problems such as association mapping of genetic traits, location of recombination hotspots, and identification of SNP sites. Moreover, algorithms that create explicit networks form a complement, or an alternative, to methods based on the more commonly used numerical, statistical, measures that less directly reflect the underlying genealogy.

1.2 Fundamental Definitions

The atomic objects of concern in this book are *individuals* in the context of population genetics, or *species* in the context of phylogenetics, or *molecular sequences* in the context of molecular evolution. Sometimes the particular biological context affects the mathematical models and the algorithmic problems that are defined on that model. However, many of the mathematical and algorithmic results we discuss in this book apply to all the biological models. We want to be as general as possible and so we will use a generic term for the objects of interest.

Definition In this book we will use the term *taxa* for the objects of interest, and *taxon* for an individual object.

Definition A *character* or *trait* is a discrete property or characteristic of a taxon. By "discrete", we mean that there is a finite number of *states* that a character can take on[4].

For example, if the taxon of interest is a human, the gender (male or female) of the taxon is a *binary character* taking on one of two possible states. As another example, the nucleotide (A, T, C, or G) present at a particular *site* of a DNA sequence is a *four-state* character. The character would be binary if we only record whether the nucleotide at that site is a purine (A or G) or a pyrimidine (C or G).

Note that the meaning of the word "character" in the context of evolutionary biology is different from the colloquial use of the word. In normal use, a character is a letter or a symbol in an alphabet, but in evolutionary biology a character is a trait of an individual. To confuse matters even more, a site in a DNA sequence can be considered as a character in the sense of evolutionary biology, but the four possible states of that character are characters in the colloquial sense of the word, i.e, in the four-letter DNA alphabet.

1.2.1 Mutation, Infinite Sites, and Binary Sequences

Definition A *point mutation* (or simply, mutation) at a single site is a change of state at that site which is independant of changes at any other site.

Note that a change of state due to a point mutation is distinct from a change of state due to recombination.

Genealogical networks represent the derivation of extant sequences which change due to both recombination and mutation. If unrestricted, mutation event alone (without recombination) can derive any set of sequences, but that derivation would not likely be biologically plausible. Therefore, we need a model of the mutations that are permitted in population sequence data.

[4] There are also continuous characters where the number of states is not finite, but these are not of concern in this book.

The most commonly used mutation model in population genetics is the *infinite sites model* where any site (in the sample) can mutate at most once in the entire studied history of the sequences. The theoretical justification is that the studied history of a sample covers a relatively short time and mutations occur at random positions, so a mutation at a given site is a low frequency event. The probability that a mutation occurs twice at a site is so low that multiple mutations can be ignored[5].

The infinite sites model implies that each site in any of the studied sequences can take on only two states, the *ancestral* and the *derived* states, and hence the sequences we observe are *binary* sequences. The strongest current validation of the binary sequence model, and of the infinite sites model, comes from DNA sequence data where each site is a *Single Nucleotide Polymorphism* (*SNP*) site, i.e., a site where only two of the four possible nucleotides appear in the population (with a frequency above some minimum threshold) [23, 93]. In humans, and other well-studied organisms [59], millions of SNP sites have been found and cataloged, most prominently by the International HapMap Project [29, 30], the Human Genome Diversity Project [118], and the One Thousand Genomes Project [31].

Note that the assumption of binary sequences is not the assumption that the DNA alphabet has been reduced from four letters to two. The DNA alphabet contains all four letters, but in the set of DNA sequences found in a population it is rare to observe more than two different letters (above a low frequency) at any given site. Any of the $\binom{4}{2} = 6$ possible pairs of differing letters can appear at any site, and many of those pairs will appear in the sequence. But because at most two different letters are observed at any site, we can code each site as a binary character and use the alphabet of $0, 1$ to represent the resulting sequence. See Figure 2.2 (on page 24) in Chapter 2.

The binary sequence assumption is also supported by morphological data in phylogenetics (representing the evolutionary history of species). There, a morphological character may be a "complex trait" caused by a succession of many uncharacterized molecular mutations. Since the complex trait requires several mutations, the probability is low that the trait will have evolved independently in different species, particularly in closely related species. Therefore, a complex trait that is common to several species is generally thought to have arisen once, in a species that is ancestral to all of the species containing that trait. However, not all complex traits are believed to obey this model, and it also believed that "convergent" or "parallel" independent evolution of highly valuable traits, such as flight or vision, has occurred.

[5] The origin of the term "infinite sites" comes from the view of a genome as having a huge (essentially infinite) number of sites so that each successive mutation, occurring at a random position in the genome, occurs at a site where no mutation has occurred before. It follows that a mutation at any given site can occur at most once.

1.2.2 The observed data

The data for most of the problems of interest consists of a set of taxa and a set of binary characters, together with information on the state of each character for each taxon. This data is usually presented in the form of a *matrix* M whose rows represent taxa, and whose columns represent characters. Each cell (f, c) of M specifies the state of character c for taxon f. For example, see Figure 2.1 on page 22 showing a matrix M with five taxa and five characters.

When talking about the matrix M we will use the terms "taxon" and "row" interchangeably, and the terms "character", "column" and "site" interchangeably, choosing whichever term is most informative for the context. Further, the ordered entries in a row f of M can be considered to form a *sequence*, and so we have

Definition The *sequence S_f for taxon f*, or the *sequence for f* is the ordered sequence formed from the entries in row f of matrix M.

Given this definition, we will also consider M to be a *set* of sequences, as well as a *matrix* representing that set of sequences. Context will often determine whether M is a set or a matrix.

1.3 A Few Graph Theoretic Definitions

The principle combinatorial objects that we deal with in this book are graphs, and so we state a few basic definitions and facts about the kinds of graphs we will encounter.

Definition An *undirected graph* $G = (V, E)$ is a combinatorial object consisting of a set of *nodes* (also called *vertices*) V, and a multi-set of *edges* E. Each edge in E is specified by an *unordered* pair of nodes from V.

Definition For an edge $e = (u, v)$ in E, nodes u and v are called the *endpoints* of edge e.

For example, the undirected graph in Figure 1.3 panel a) has node set $V = \{a, b, c, d, e\}$ and edge set $E = \{(a, b), (a, c), (a, e), (b, d), (c, d), (c, e)\}$.

The definition of an undirected graph allows an edge whose two endpoints are the same, creating a *self-loop*. The definition of an undirected graph also allows multiple copies of the same edge, creating *parallel* edges. This is the reason that E is formally a multi-set rather than a set. In this book, we do not need graphs with self-loops or parallel edges.

Definition An undirected graph without self-loop or parallel edges is called a *simple graph*. In that case, E is a set of edges.

Definition A *directed graph* $G = (V, E)$ is defined by a set of nodes V, and a multi-set of *directed edges* E, where each directed edge is specified by an *ordered* pair of nodes. By convention, the directed edge $e = (u, v)$ is directed from node u to node v. Node u is called the *tail* of e and node v is called the *head* of e. See Figure 1.3 b).

A directed graph without self-loops or parallel edges is called a *simple directed graph*. In this book we will not need directed graphs with self-loops or parallel edges.

In this book we will refer to a simple graph as a "graph" for short, and a simple directed graph as a "directed graph" for short. However, we will sometimes explicitly use "undirected graph" to emphasize the undirected nature of the edges. When the graph is directed, we will explicitly refer to it as a "directed graph".

Definition If G is a directed graph, the *underlying undirected graph* of G is the graph formed by ignoring the directions on the edges of G. That is, each ordered pair of nodes that defines an ordered edge in G is now considered as an unordered pair of nodes.

Definition For any node v in an undirected graph, the *degree* of v is the number of edges of that touch v, i.e., the number of edges where v is one of the endpoints. For a node v in a directed, the *in-degree* of v is the number of edges directed into v, i.e., where v is the head of the edge; the *out-degree* of v is the number of edges directed out of v, i.e., where v is the tail of the edge.

Definition An *undirected path* from a node v_1 to a node v_k in an undirected graph $G = (V, E)$ is specified by an ordered list of nodes $v_1, v_2, \ldots, v_k$, such that for every i from 1 to $k - 1$, the node pair (v_i, v_{i+1}) is an edge in E.

Definition A *directed path* from a node v_1 to a node v_k in a directed graph $G = (V, E)$ is specified by an ordered list of nodes $v_1, v_2, \ldots, v_k$, such that for every i from 1 to $k - 1$, the ordered node pair (v_i, v_{i+1}) is an edge in E, i.e., an edge directed from v_i to v_{i+1}.

Definition An undirected graph G is *connected* if for ever pair of nodes u, v in G there is a path between u and v in G. A directed graph G is connected if the underlying undirected graph of G is connected. A directed graph G is *biconnected* if for every ordered pair of nodes u, v, there is a directed path in G from u to v.

Definition An *undirected cycle* in an undirected graph G is an undirected path which starts and ends at the same node. A *directed cycle* in an directed graph G is an directed path which starts and ends at the same node.

1.3.1 DAGs and Trees: The most important graphs in this book

Now we can define the most important types of graphs discussed in this book, and some properties of those graphs.