Introduction

> In the long history of humankind (and animal kind, too) those who learned to collaborate and improvise most effectively have prevailed.
>
> *Charles Darwin*

> Knowing is not enough; we must apply. Willing is not enough; we must do.
>
> *Johann Wolfgang von Goethe*

This book is about the application of algebraic topology to the problem of organizing and describing biological data. The problems this book studies are of recent origin. For much of its history biology was a predominantly descriptive science with comparatively little interaction with mathematics. Explanations of mechanism took place at the level of entire organisms or cells. But over the last century, the development of molecular biology has transformed the field so that it is now data intensive and marked by increasing reliance on mathematics.

This shift began with the discovery of the elemental constituents and rules that govern biological systems at the molecular level. Early highlights included the determination of the structure of DNA, RNA, and proteins, the deduction of some of the processes of information transmission within the cell, and the identification of specific molecular mechanisms underlying particular biological processes.

For a long time, small amounts of data were hard-won in the laboratory; for example, many researchers were focused on elucidating the biological mechanisms of individual genes, the sequences of DNA that are translated into RNA and produce a functional product such as a protein. However, towards the end of the twentieth century, the rate of data production accelerated very rapidly and it became possible to study all the genes of a cell (the entire genome) at once. The publication of the first draft of the human genome [513] in 2001 was a milestone in this revolution, heralding transformation in almost every realm of biology.

An incomplete sampling of the subsequent progress on fundamental problems includes the enumeration of genomic variations in thousands of individuals [122], detailed molecular characterization of thousands of cancers [343], single cell characterization of tumors [401], study of developmental processes [504], and the elucidation of the three dimensional structure of DNA in the nucleus of cells [138, 330].

Mathematics has played a key role in the development of modern molecular biology. The amazing progress in data collection depended in part on the development of mathematical algorithms that supported the assembly of raw DNA sequencing information and enabled the search for genes in the sequences. The development of and continued research on these algorithms is a fascinating and deep story, but it is not the focus of our inquiry here. Rather, we will study mathematical tools for determining the structure of biological processes and mechanisms from the data.

Analyzing biological data is a difficult problem. There is a large amount of data, and it is particularly challenging to work with: high-throughput genomic and transcriptomic data typically resides in very high-dimensional spaces (e.g., on the order of the number of genes in the organism, which can be in the tens of thousands), is frequently extremely noisy, and often reflects poorly understood systematic errors. For example, genetically similar organisms or cells can display different molecular profiles (e.g., present different epigenetic states, express different genes) leading to markedly different experimental measurements.

In short, modern biology has become a data rich discipline, dependent on sophisticated mathematical techniques for both the production of experimental data and its interpretation. In this way, it exhibits kinship with modern physics. But in contrast to the situation in physics, the mathematical models we have to understand genomic processes are in general less descriptive and provide fewer conceptual benefits than the models of physics. One problem is that the immense complexity of fundamental biological systems means that we simply lack good theoretical frameworks to describe them. For example, the enormously complicated cycle controlling gene expression is still not completely understood. Even our knowledge of the basic objects of study is incomplete; we hear almost daily that a new noncoding gene has been identified or that a novel viral species has been associated to a newly reported disease.

The point of departure for this book is a concrete manifestation of this lack of models: to date there has been no real analogue in biology for the role of geometry in physics. Geometry is at the heart of modern physics. This is no surprise; in a sense, modern geometry was invented to describe physical systems. Calculus was developed in order to describe the acceleration of moving bodies. Einstein's theory of general relativity can be succinctly summarized as the contention that gravity curves spacetime, which can be precisely and concisely expressed in the

language of differential geometry. In stark contrast, biological data does not naturally appear to have the same kind of rich geometric structure. Typically, all one has is a collection of data points and various choices of a way of measuring the distance between them. Even if there was geometry present, it might be hard to see through the noise.

Our central dogma in this book is that although biological data might not possess the rigid geometric structure that arises in physics, it nonetheless has meaningful coarser geometry; we will broadly refer to this as *shape*. In some sense, this hypothesis is implicit in the standard approach for analyzing genomic data, namely dimensionality reduction and clustering. We can access the geometry of the situation through a distance function that takes as input a pair of data points and outputs a number (larger than 0) that reflects the distance between them. (Here distance is an abstract notion, not a measure of physical distance.) Dimensionality reduction refers to the process of using the distances to embed the data points (which might lie in a 10000-dimensional space) into a low-dimensional space (like the standard two dimensional Euclidean space $\mathbb{R}^2$) in such a way that distances are preserved as much as possible. Clustering refers to the process of grouping the data points into "clusters" such that points within a cluster are much closer to each other than to points in distinct clusters. Often these techniques are combined; clustering algorithms are applied to the results of dimensionality reduction, and we will sometimes refer to the combination as "clustering analysis."

Clustering genomic data has been a very successful way to detect genomic relationships with clinical consequences. In Figure 0.1, there is a representative example of a clustering analysis of mRNA expression data from pancreatic tumors. The data, obtained from samples from 147 patients, consists of vectors of numbers representing the expression levels for various genes. The distance between these expression vectors is roughly speaking a measure of similarity; tumor samples with similar expression profiles are close together. Then the data naturally breaks up into three clusters of points, as indicated in the plot on the left side of Figure 0.1. Each column represents the expression vector of a particular tumor sample; each row represents a particular gene. A point in the square thus encodes the level of expression of a gene in a sample – red means highly expressed, blue suppressed.

It is clear from the picture that points within a cluster have similar expression profiles, but more importantly, these clusters are clinically significant – which cluster a tumor sample is in predicts survival rates. Figure 0.2 graphs the survival curves for the different clusters; squamous pancreatic adenocarnicomas (cluster 2) have noticeably worse survival trajectories. That is, understanding the shape of the expression data as captured by clustering allows us to predict the likely progression of the cancer.

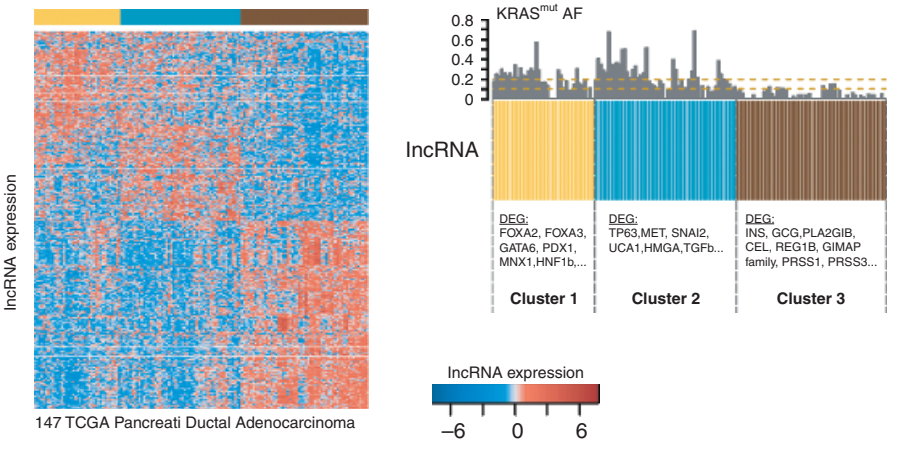

Figure 0.1 Using mRNA expression from long non-coding RNA from 147 patients with pancreatic adenocarcinomas, one can observe three different clusters. Cluster 2 is associated to squamous pancreatic adenocarcinomas. Different clusters reveal molecular mechanisms common to a set of patients. Source: [17]. Reproduced from *Gut*, Luis Arnes, Zhaoqi Liu, Jiguang Wang et al., Published Online First: 10 February 2018. © 2018. With permission from BMJ Publishing Group Ltd.

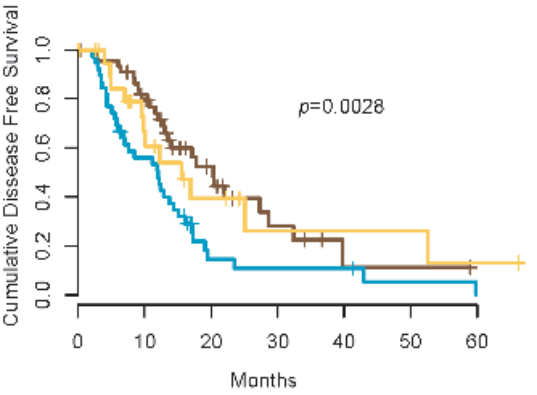

Figure 0.2 Different clusters of pancreatic adenocarcinomas have very different survival profiles. The *y*-axis represents the fraction of patients as a function of time. The colors represent the different clusters. Ideally, we would like to assess the prognosis of a patient based on molecular characteristics, and clustering patients constitutes a simple way of doing that. In addition to the clinical correlates, different clusters could reflect different molecular mechanisms that lead to the disease. Source: [17]. Reproduced from *Gut*, Luis Arnes, Zhaoqi Liu, Jiguang Wang et al., Published Online First: 10 February 2018. © 2018. With permission from BMJ Publishing Group Ltd.

More generally, dimensionality reduction and clustering methods such as PCA, MDS, spectral clustering, non-negative matrix factorization, and so forth are ubiquitous tools for analysis of genomic data. However, despite their successes, clustering algorithms capture only a very limited amount of information about shape – they are sensitive only to how many disconnected pieces a data set should be separated into. And this is often not enough – for example, there are many data sets of tumor samples where the points do not naturally separate into clusters which correlate with clinical outcomes.

Moreover, there are many other questions one can pose about the shape of a data set. For example, a natural question that arises when studying evolutionary phenomena is whether or not genomic data (for example, sequencing information from different flu viruses) can be represented by a tree structure, where the lengths of the branches correspond to distances between data points. To answer this question, an obvious approach is to attempt to determine if the points are better represented not by a tree but by a graph with loops. Such shape information cannot be extracted from clustering, and traditional dimensionality reduction algorithms tend to introduce distortions that obliterate this kind of shape.

Our aim in this book is to make the case that robust algorithms for capturing high-dimensional shape can be effective in situations where clustering fails. Specifically, we want to explain particular mathematical tools from algebraic topology that generalize clustering algorithms, giving rise to a methodology for extracting scientifically meaningful high-dimensional shape information from genomic data.

0.1 Why Algebraic Topology?

Modern algebraic topology was invented by Poincaré to provide tools for describing global properties of differential equations on surfaces. His basic insight was that the qualitative behavior of differential equations depended on the shape of the underlying surface. Algebraic topology studies qualitative and often global properties of geometric objects by constructing *algebraic invariants* of such objects.

By **geometric object**, we mean what we will refer to as a "space"; for the purposes of current discussion this means a subset of Euclidean space (e.g., the surface of a rubber band, or a sheet of paper, or a soda can). By **algebraic**, we mean something like a number or a vector space. By **invariant**, we mean something which is not changed by **smooth deformation**; stretching is allowed, but not tearing, as if we were studying things made out of soft clay. By **global**, we mean something that cannot be figured out by looking at a little piece of the object – one has to inspect the entire thing.

Let us begin with a very simple example. Suppose we want to answer the geometric question of distinguishing between two collections of non-overlapping solid blobs; given a collection of n solid blobs (referred to as the "disjoint union") and the disjoint union of k solid blobs in the plane $\mathbb{R}^2$, we want to decide if these pictures are the same or different. (See Figure 0.3 for a representative example.) An easy way to do this is to count the number of *path components* – the number of distinct pieces that cannot be connected by a path, i.e., a line drawn without removing the pencil from the page. For example, in Figure 0.3, the left hand shape has three path components and the right hand shape has five.

This count is a simple example of an algebraic invariant; it is a number, and it is not affected by smoothly deforming the blobs. Moreover, it is clearly a global quantity – just looking at a little piece of one of the blobs or even any finite subset of the blobs will not suffice to compute it. And using this count allows us to distinguish between geometric objects simply by the algebraic operation of comparing two numbers. Notice that counting path components feels very reminiscent of clustering! And as we will explain, there is a precise relationship between these procedures.

The count of path components is a fairly crude invariant of a space. But there are many more sophisticated invariants which can detect more interesting properties of the shape of a geometric object. Figure 0.4 shows a more difficult version of the question about blobs: how can we distinguish between a circle and a figure-eight?

Figure 0.3 On the left, there are three path components, on the right, five path components.

Figure 0.4 A circle (or annulus) has a single hole; a figure-eight (or union of two annuli) has two holes.

Counting path components cannot distinguish these spaces; there is a single path component in each case. However, if we count the number of "holes" or closed loops, we see that the figure-eight has two holes whereas the circle has one hole. This is another global invariant, the first Betti number, which counts the number of "holes" enclosed by circles in a geometric object. Once again, notice that smoothly stretching the circle and the figure-eight will not change the Betti number.

These examples suggest that algebraic topology provides a powerful methodology for capturing robust global properties of the shape of geometric objects and turning them into algebra. But it is a priori not clear how to use these tools to study real data! The questions we have been discussing above have used spaces that are defined as infinite sets of points, most concisely specified by equations. This observation raises two important issues. First, one might worry that describing spaces in this way does not seem to be algorithmically tractable. Second, the data sets of biology are likely to be finite sets of isolated points – how can we associate a continuous space to a finite set?

0.2 Combinatorial Algebraic Topology

Conveniently, there is a long tradition in algebraic topology of studying *combinatorial models* of geometric objects. By *combinatorial*, we mean a description of a space using only discrete data. Such models are well suited to algorithmic computation. The most important kind of combinatorial model for the approach discussed in this book is the *simplicial complex*. We will give a precise definition of a simplicial complex in Section 1.8, but roughly speaking a simplicial complex should be thought of as a geometric object specified by gluing together a collection of points, line segments, triangles, and higher dimensional analogues called simplices. Simplicial complexes represent spaces up to continuous deformation; they are a satisfactory representation for computing topological invariants.

Figure 0.5 presents examples of the standard pieces (called simplices) that are glued together to form the space represented by a simplicial complex. A k-dimensional simplex has *faces* which are $(k-1)$-dimensional simplices; a simplicial

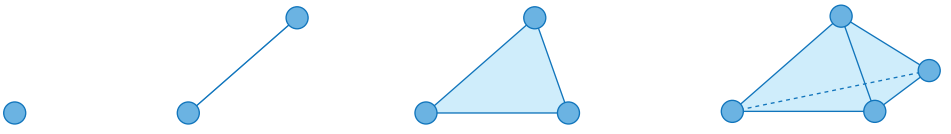

Figure 0.5 Simplicial complexes model spaces made by gluing together standard triangular pieces, called simplices; here we illustrate the 0-, 1-, 2-, and 3-dimensional simplices.

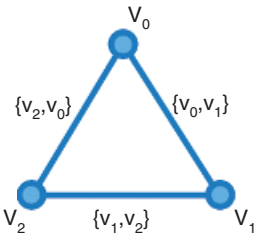

Figure 0.6 A simplicial model of the circle is given by gluing three 1-simplices together at their endpoints.

complex is made by gluing together standard simplices along their faces. For example, the faces of a 1-simplex are the two endpoints. The faces of a 2-simplex are the three edges of the triangle.

To describe a simplicial complex, one simply specifies the number of 0-simplices, 1-simplices, etc. as well as instructions for gluing them together. For example, a simplicial complex consisting only of 0-simplices and 1-simplices is specified by a collection of edges and instructions for attaching them at their endpoints – this is precisely the data of a graph, with 0-simplices the vertices and 1-simplices the edges. That is, a simplicial complex is precisely a higher dimensional generalization of a graph.

For example, consider the complex in Figure 0.6. We have 0-simplices $\{v_0, v_1, v_2\}$ and the 1-simplices $\{\{v_0, v_1\}, \{v_1, v_2\}, \{v_2, v_0\}\}$ where, for example, the 1-simplex $\{v_0, v_1\}$ has faces v_0 and v_1, and is thought of as a line segment connecting the vertices. We think of this complex as *representing* a circle; we are working with spaces up to continuous deformation, and a triangle can be stretched out into a circle.

Notice that there are other natural ways to represent a circle: one could use the "square" specified by the 0-simplices $\{v_0, v_1, v_2, v_3\}$ and the 1-simplices $\{\{v_0, v_1\}, \{v_1, v_2\}, \{v_2, v_3\}, \{v_3, v_0\}\}$.

Figure 0.7 shows how to produce a simplicial model of a solid disk (the circle plus its interior): we could take our first model of the circle above and add the 2-simplex with faces the 1-simplices; we can uniquely specify this 2-simplex as $\{v_0, v_1, v_2\}$. Here the 0-simplices and 1-simplices of the circle specify the *boundary* of the disk; the 2-simplex describes how the *interior* is glued to the boundary. An example of a more complicated simplicial complex is shown below in Figure 0.8; this represents a hollow ball with a circle attached to it (at the vertex v_4) and a line attached to the circle (at the vertex v_2).

The real payoff from working with simplicial complexes is that algebraic invariants of the spaces they represent can be algorithmically computed directly from the combinatorial description. The prototypical example of an algebraic invariant associated to a simplicial complex is the *Euler characteristic*. Suppose that we

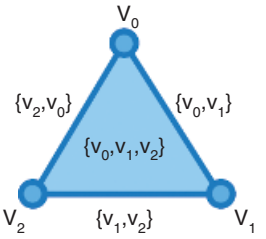

Figure 0.7 A simplicial model of the solid disk is given by gluing three 1-simplices together at their endpoints and gluing a 2-simplex to them at its faces.

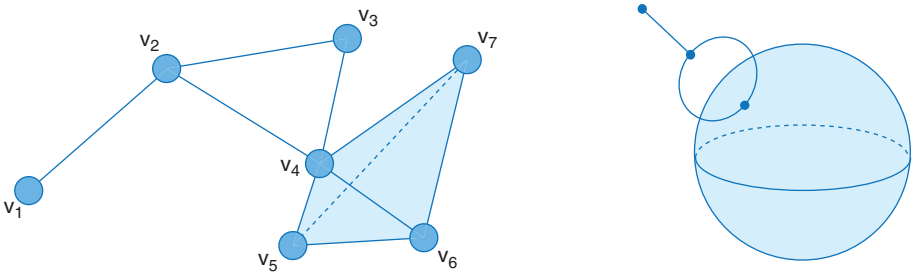

Figure 0.8 The simplicial complex on the left is made by gluing together the standard simplices. Combinatorially, we would write this as having 0-simplices $\{v_1, v_2, v_3, v_4, v_5, v_6, v_7\}$, 1-simplices $\{\{v_1, v_2\}, \{v_2, v_3\}, \{v_2, v_4\}, \{v_3, v_4\}, \{v_4, v_5\}, \{v_4, v_6\}, \{v_4, v_7\}, \{v_5, v_6\}, \{v_5, v_7\}, \{v_6, v_7\}\}$, and 2-simplices $\{\{v_4, v_5, v_6\}, \{v_5, v_6, v_7\}, \{v_4, v_6, v_7\}, \{v_4, v_5, v_7\}\}$. On the right is a space represented by this complex.

have a simplicial complex with V vertices, E 1-simplices, and F 2-simplices and no higher simplices. Then the Euler characteristic is $V - E + F$. In general, the Euler characteristic of a simplicial complex is the alternating sum of the numbers of k-simplices.

For example, the Euler characteristic of a point is clearly 1. The Euler characteristic of a simplicial complex consisting of a single 1-simplex and its two endpoints is $2-1 = 1$. Next, consider the simplicial complex modeling the circle from the discussion above – this is a loop formed by the three vertices and three line segments. This complex has Euler characteristic $3 - 3 = 0$. If we take the model of the circle given by the "square," this also has Euler characteristic $4 - 4 = 0$, and in general any such model of a circle will have n vertices and n 1-simplices and hence Euler characteristic 0.

On the other hand, the Euler characteristic of the disk given by filling in the triangle with a 2-simplex is $3 - 3 + 1 = 1$. Notice that the Euler characteristic of the filled triangle is the same as the Euler characteristic of a point; the Euler

characteristic is a topological invariant, as it is insensitive to smoothly crushing the triangle to the central point. Comparing the results for the triangle and the filled triangle, we observe that the Euler characteristic is detecting a topological property, namely that the loop has a hole in the middle.

We can also compute the path components of a space represented by a simplicial complex directly from the complex; this turns out to reduce to a standard problem in graph theory. More generally, one can compute many algebraic invariants directly from simplicial complexes – for instance, the problem of counting holes (i.e., the Betti numbers) can be transformed into an elementary problem in linear algebra, as we shall see in Section 1.10. In summary, provided that we can represent our data using an appropriate simplicial complex, we can apply the computational tools of algebraic topology.

0.3 Topological Data Analysis (TDA)

The kind of biological data we will work with is typically presented as a finite set of points equipped with some kind of distance or dissimilarity measure between the points; a mathematical model of this situation is a *finite metric space*, which is a set X of points equipped with a distance function ∂_X satisfying a few simple axioms. The central question is: given data presented as a finite metric space, how can we robustly produce a simplicial complex such that the algebraic invariants of the simplicial complex reflect the shape of the data? Often, we hypothesize that these points are samples from a probability distribution on some geometric object; Figure 0.9 gives an idealized picture of this situation.

Consideration of clustering guides us to an answer. To explain, we need to make the connection between clustering and components precise, via *single-linkage clustering*, which works as follows.

1. Fix a scale parameter ϵ.
1. Assign two points x and y to the same cluster if they are connected by a path of points (for some k)

$$x = x_0, x_1, x_2, \ldots x_{k-1}, x_k = y$$

such that each point x_i is within a distance ϵ of x_{i+1}.

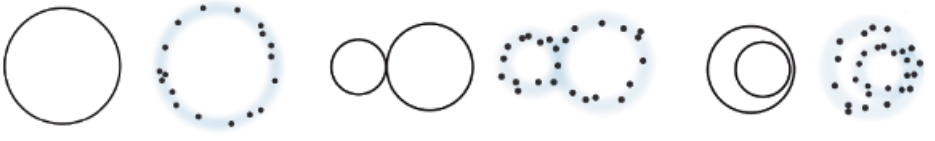

Figure 0.9 On the left, the underlying geometric "ground truth." On the right, finite samples from which we seek to recover the invariants of the circle, figure-eight, and nested circles.

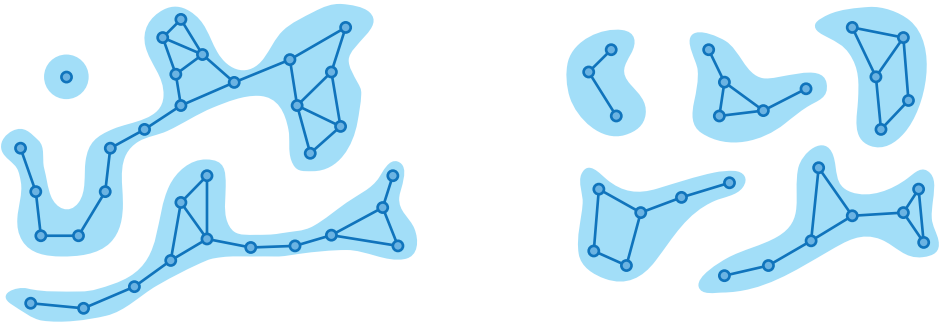

Figure 0.10 The clusters are the path components of the simplicial complexes.

We can interpret single-linkage clustering in terms of simplicial complexes: form the simplicial complex whose vertices are the data points and whose 1-simplices connect points x and y if they are less than a distance ϵ apart. Now the single-linkage clusters are precisely the path components of this simplicial complex; see Figure 0.10 for an illustration. But we can go even further – namely, we can add higher dimensional simplices when groups of points are close in some way. For instance, we could add a 2-simplex for every triple of points $\{x, y, z\}$ such that each pair $\{x, y\}$, $\{x, z\}$, and $\{y, z\}$ has distance less than ϵ. We then hope that the topological invariants of the resulting simplicial complex are capturing qualitative information about the shape of the data set.

This procedure has some attractive properties. Sufficiently small perturbations of the data typically result in small perturbations of the resulting simplicial complex that do not change algebraic invariants. Moreover, the simplicial complex constructed in this fashion reflects the sensible hypothesis that small measured distances between data points are likely to be accurate, but large distances are probably not accurate and should instead be estimated in terms of small distances. So intuitively speaking, it seems plausible that such topological invariants will be robust against certain kinds of noise and corruption, and will reflect real geometric structure of the data.

However, choosing ϵ correctly is difficult; this requires some knowledge of the feature scale of the data. It is illuminating to reflect on what happens to these simplicial complexes as ϵ increases; see Figure 0.11. When ϵ is very small, there are just discrete points (panel A). When ϵ is larger, the resulting simplicial complex has interesting geometric structure (panels C and D). And when ϵ is very large, everything is connected and there is no information recovered at all (panel E).

In the example above, it is not clear what the "correct" value of ϵ is, as the underlying topology is not evident. The best we can say is that there is a wide range of values for ϵ in which there is non-trivial topology. In simple cases, however, we

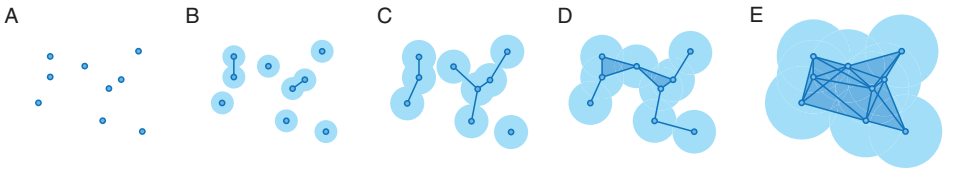

Figure 0.11 As ϵ grows, more and more simplices are added to the simplicial complex.

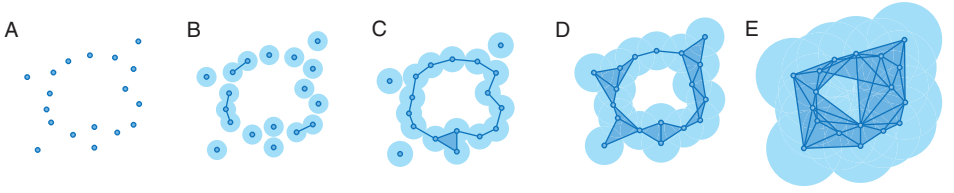

Figure 0.12 As ϵ grows, topological features appear. In panels C and D, the circle can be detected.

might hope to extract more precise topological hypotheses; we illustrate how this might work in Figure 0.12.

When ϵ is small, again the result is just discrete points (panel A). As ϵ grows, adjacent points begin to link up (panel B). But there is a wide range in which ϵ results in adjacent points along the circle being connected without connecting points across the circle (panels C and D); this is an illustration of the importance of privileging "short" distances over "long" ones. One way of looking at the situation is to observe that for these ϵ, distances between points that are less than ϵ accurately reflect distances along the circle. When ϵ is large enough, connections across the circle "short-circuit" the complex (panel E), and we eventually again obtain a completely connected complex.

As both of the preceding examples make clear, it is a priori very difficult to guess what the correct feature scale should be. There might be multiple scales at which we expect to see meaningful topological features, or it might even be the case that no single scale correctly encodes the salient features. A basic philosophy underlying topological data analysis is that scale issues should be handled simply by encoding the complexes for all ϵ simultaneously and keeping track of how they change as ϵ changes. This leads to a series of new algebraic invariants, which reflect the *persistence* of topological features across scales. By using these invariants, topological data analysis provides tools for robustly describing multiscale shape information of data.

In recent years, there has been an explosion of work in this area; however, many interesting problems remain to be solved. For instance, there are still many

questions about the relationship between statistical practice and the invariants of topological data analysis. Nonetheless, part of our motivation for writing this book is that already the field is sufficiently mature for there to be many interesting applications to biological data. With this in mind, we now explain why topological data analysis is a potentially very useful tool to analyze biological data. We begin by explaining the kinds of biological problems that we will focus on in this book.

0.4 Genetics and Genomics

We will focus on biological questions arising from the perspectives of modern genetics (the study of genes, the fundamental units of heredity) and genomics (the study of genomes, the collections of all genes in an organism). These questions have been chosen to illustrate how topological data analysis can be used to address biological problems. Genetics and genomics are particularly amenable to the application of topological methods: there is a great need for mathematical tools to study the shape of large amounts of large scale experimental data, and the standard methods in use are comparatively crude.

At a high level, most of the problems in genetics can be posed in a simple fashion. There are two "spaces" of interest, the space of genotypes (the set of possible genomes) and the space of phenotypes (the set of observable characteristics of an organism that could occur in a particular environment); scientific questions are typically about describing and understanding a function that maps genotype into phenotype (e.g., see Figure 0.13). Such a function specifies which genetic alterations lead to a particular phenotype. Conversely, the function also determines the

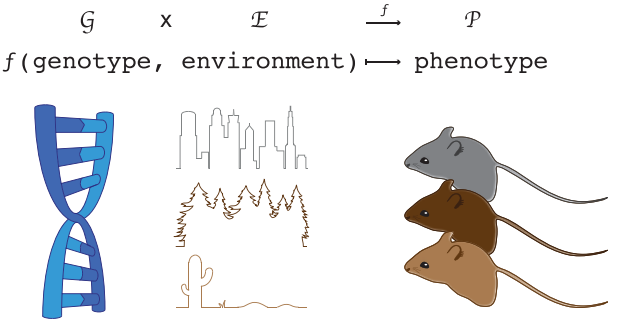

Figure 0.13 Many problems in genetics can be posed as the study of functions from the space of genes and genomes (the genotype) and the environment to the observable characteristics (the phenotype). Variation in genes and genomes between different organisms causes changes in observable characteristics, such as protein structure, protein function, disease survival, and many other potential phenotypes.

most interesting phenotypes to look at when studying variations in a particular gene. For example, cancer genetics studies the impact of mutations in cancer cells (genotypes) on clinical manifestations of cancer, notably on tumor growth, tissue invasion, and metastasis (phenotypes). To understand this relationship, one studies the molecular mechanisms of this association: how a mutation changes a protein, how this change affects the cell, and how such changes lead to the observed phenotypes.

At a very high level, the genome can be understood as a long word whose letters are the four nucleotide bases, denoted (A, C, G, and T or U, in the case of RNA). The length of this word varies dramatically across different organisms. The shortest, called viroids, are a few hundred bases. Humans have roughly three billion bases. And plant genomes can be two orders of magnitude larger (e.g., the genome of *Paris japonica*, a rare and beautiful plant from alpine regions in Japan, has a genome of 150 billion bases). The situation is further complicated by the fact that in multicellular organisms, such as humans, different cells will have similar but not necessarily identical genomes. Mathematically, different organisms can be regarded as producing distinct points in the genotype space, and so can different cells from a single organism.

However, the most interesting sources of variation come from the fact that genomes are not stable objects; they change over time. Specifically, errors occur when the genome of an organism is copied to produce offspring. The simplest types of mistakes are *point mutations*, where at a particular place in the genome one base is replaced with another one. However, more drastic changes can occur; sections of the genome can be lost or duplicated, or there can be more wholesale scrambling. In Figure 0.14 we show the typical order of magnitude of the size of the genome of different organisms along with the mutation rates (i.e., the probability of a point mutation at a particular spot per replication). Genome sizes and mutation rates vary by orders of magnitude; organisms with shorter genomes tend to be prone to mutations. A pervasive and more complicated phenomenon is that different organisms can exchange genomic information, resulting in new genomes which shuffle together the original genetic information. These processes produce clouds of points in the genome space; the problem is then to understand the relationship between these point clouds and resulting phenotypic changes.

However, the phenotype space is harder to specify and often more complex than the genotype space. Examples of phenotypic characteristics include the expression of different mRNAs, the expression of proteins, the shapes of these proteins, the shape of the cell, the ability to grow and replicate, the susceptibility to different stimuli, the ability to respond to an infection, and the size and weight of a multicellular organism. Obviously we do not expect an exhaustive enumeration of scientifically important observable characteristics. Instead, different areas of biology focus on specific choices of salient phenotype; for example, in evolutionary

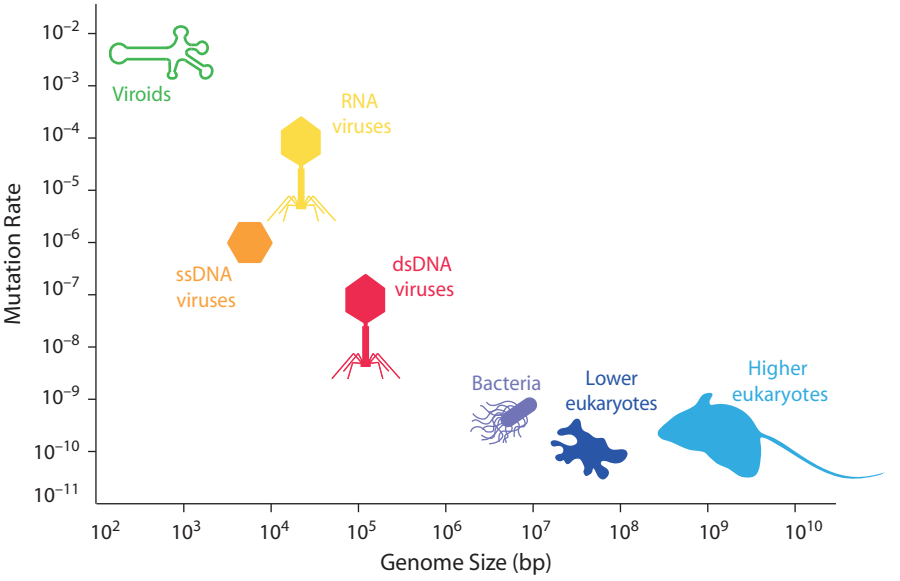

Figure 0.14 The size of genomes varies by many orders of magnitude. Viroids are small (a few hundred bases) sequences of free RNA that can infect plants. The genomes of RNA viruses (like influenza) are usually around 10,000 bases, the genomes of DNA viruses are typically 100,000 bases, and the genomes of bacteria can be millions of bases. Some plant genomes can reach 100 billion bases (*Paris japonica*). There is a fascinating relationship between the size of the genome and the number of mistakes per replication (mutation rate), represented here in the *y*-axis.

biology we might be interested in fitness or the ability to proliferate in some particular environment. In the context of tumors, proliferation, invasion of new tissues, and survival rate are all interesting phenotypes to study.

Finally, the environment is also an important factor determining when genetic variation will cause changes in the phenotype. Many genes are only expressed in certain circumstances, and beneficial alterations in one environment could be detrimental in another one. As a first approximation to reduce the complexity of the problem, it is common to fix or reduce the number of environmental factors to a few conditions that are suspected to be germane to the phenotype under study.

0.5 Why Is Topological Data Analysis Useful in Genomics?

Our contention is that topological data analysis provides novel and effective tools to attack the problem of inferring the relationship between genotypic events and changes in phenotype. For example, understanding the shape of the data in genome space reveals the way that certain phenotypical changes arise. To support our claim,

in this book we describe a number of biological problems where the methods of topological data analysis reveal new and interesting phenomena. In each case, biological data is presented as a finite collection of points equipped with a distance, i.e., a finite metric space. Topological invariants of associated simplicial complexes then turn out to encode biologically relevant quantities. Here, we describe three illustrative examples (see Figure 0.15).

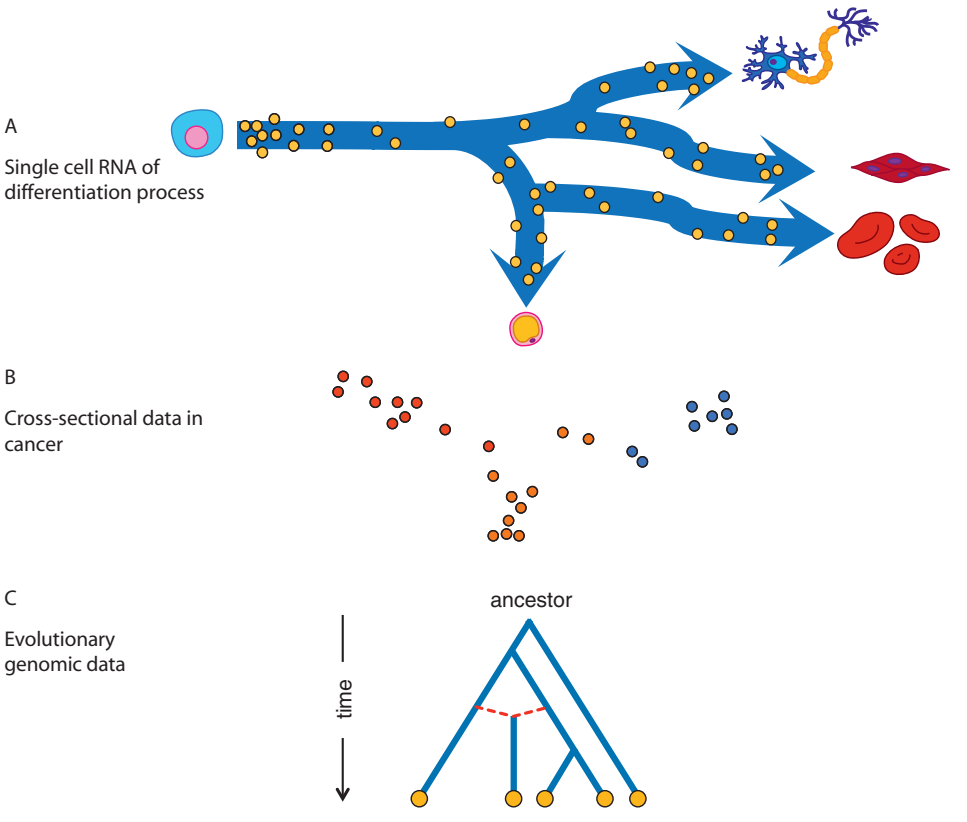

Figure 0.15 Examples of biological point cloud data: (A) Starting from stem cells, different cell types arise from a process of differentiation over time. Single cell approaches provide information about differentiation, where each point corresponds to a particular cell. Important questions include characterizing distinct subpopulations/expression programs/specific surface markers, and determining how cells decide their fate. (B) Each point represents the tumor of a patient. Questions in this space concern the classification of patients according to their molecular profile, association of location in this space with survival, determination of mechanisms of drug resistance, and the identification of specific pathways implicated in tumor progression. (C) Each dot (tree leaf) represents a genome. Traditionally, we expect evolutionary processes to be described as trees. But there are many examples of phenomena (e.g., recombinations) that do not fit into this framework. This raises the question of how to describe the relationship between genomes.

Our first example concerns the process of differentiation (panel A). A baby animal begins from a single cell that divides and differentiates to generate an incredibly complicated collection of organs, tissues, and cell types. Differentiation is usually represented as a branching process, with a root, the stem cell, giving rise to descendant cells of many different types. All of our cells share a common genome: a particular cell type is characterized by the *expression* of specific genes, i.e., by the amounts of messenger RNA (mRNA) or protein generated from each gene. Transcription, the generation of RNA copies from DNA, follows a carefully orchestrated program in which certain genes are turned on in consonance with other genes. This transcription program is regulated by proteins that control the expression of multiple genes; the regulation ensures that the right amounts of RNA are produced at the right times. Cells of similar type have related transcriptional programs and thus similar gene expression profiles – comparing the expression of genes of individual cells sampled along the process of differentiation can reveal the specific mechanisms that determine what type of cell will arise.

Until recently, studying the process of differentiation was complicated by the fact that experimental techniques commingled genomic information from many cells, each potentially in a different stage of differentiation. However, single cell expression technologies now allow the measurement of the transcriptional state of single cells throughout the differentiation process. The transcriptional state of a cell can be described by a vector $(e_1, e_2, \ldots, e_G)$, where e_i is a measure of the amount of mRNA produced from gene i and G is the total number of genes measured. Given this data, we wish to characterize different cell states and types and infer the trajectory of the differentiation process. This problem can be formulated as reconstructing low-dimensional geometric structure from a sample of points (cells sequenced) in a high-dimensional ambient space (with dimension given by the number of genes). Both the Euclidean distance and the correlation between expression vectors can be used to provide metrics on transcription vectors, and clustering using these metrics has been applied to great effect. Framed in this fashion, the problem of analyzing differentiation using single cell data is clearly a potential application area for the tools of topological data analysis.

Our second example focuses on cancer (panel B). A cancerous tumor is the result of the accumulation of mutations that lead to uncontrolled cell growth; for example, mutations that alter the cell division cycle by reducing apoptosis (cell death), enhancing blood supply, and increasing generation or responsiveness to growth-promoting signals. Tissue samples from tumors in patients can be sequenced to identify these mutations and to try to determine how expression differs between tumors and normal tissue. While each individual tumor is the result of specific genomic alterations, almost all currently available therapies are generic. Moreover,

it is often unclear why some tumors are cured by treatment whereas others progress. The goal of *precision medicine* is to provide doctors with guidance enabling the deployment of therapies tailored to a particular patient's tumors.

Given sequencing data from a variety of tumors, one can study the relationships between these tumors in the hope of finding improved tumor classifications, discovering specific molecular mechanisms of progression and drug resistance, and eventually providing specific therapeutic options based on the tumor's molecular profile. Sequencing data taken from the same tumor at different times gives insight into tumor progression and development. Traditional computational approaches cluster the data and use the clusters to try to characterize the spectrum of alterations, pathways, and the clinical characteristics (e.g., survival). However, often the structure of the data does not support unambiguous division into clusters; in this case, the task of grouping patients is very difficult, as the number of clusters becomes a matter of opinion and many of the samples remain unclassified. Better tools for understanding and characterizing the shape of the tumor data have the potential to provide valuable information about clinical relationships. As we explain, sophisticated geometric models of the space of tumors as well as simplicial complexes associated to the metric space of sequencing data reveal biologically meaningful structure invisible to clustering algorithms.

Our final example has to do with evolution (panel C). Darwin first proposed the phylogenetic tree as a means to represent the evolution of phenotypic attributes. Since then, methods in molecular phylogenetics have been developed to characterize evolutionary relationships between species. These approaches generally assume that genomic information is solely passed from parents to children.

However, it has long been known that more complex modes of genetic exchange can occur, including lateral gene transfer in bacteria, recombination and reassortment in viruses, viral integration in eukaryotes, and fusion of genomes of symbiotic species. These "horizontal inheritance" phenomena can cause serious concerns about the reliability of inference of evolutionary relationships.

For example, traditional phylogenetic classifications of microorganisms have relied on evolutionary relationships inferred from 16S ribosomal RNA, a highly conserved genomic region between bacteria and archea species. However, as this region accounts for under 1% of the complete genome in most species, the vast majority of genetic information is ignored. Since horizontal inheritance is pervasive, the remaining 99% of genes might tell a very different evolutionary story. This problem becomes acute in viruses that lack 16S or other universal genes. Such challenges underscore the need for approaches to describe the shape of evolutionary processes in a more general way, free from the constraints of the tree representation. As a first step, it would be very useful to have criteria to determine when tree

representations are inadequate. The rapidly growing number of sequenced microbial genomes provides fertile ground for developing and testing new approaches to quantifying both vertical and horizontal evolutionary processes.

While recent developments in phylogenetic networks provide ways to identify instances of non-tree-like events, the field does not have a widely accepted framework to visualize and quantify the frequency, scale, and significance of horizontal evolution. Although phylogenetic trees can be visually complex, from a topological standpoint they are very simple mathematical objects: a tree is a simplicial complex with only 0-simplices and 1-simplices that contains no loops. In contrast, simple kinds of horizontal evolution can be represented by the presence of loops. Thus, computing Betti numbers of the simplicial complexes produced from sequencing data can detect horizontal evolution.

0.6 What Is in This Book?

This book is aimed at two distinct audiences: quantitative biologists interested in applying new mathematical tools to the study of genomics, and mathematicians and computer scientists interested in understanding geometric problems that arise in modern genetics and genomics. As a consequence, we have written neither a traditional mathematics textbook nor a standard biology textbook.

In the first part of the book, we begin by giving a rapid but comprehensive review of the mathematical background for topological data analysis (TDA). We state definitions and theorems, and provide many examples, but do not give proofs; our goal is to provide context for understanding the TDA framework and also to provide detailed references for the reader interested in achieving a deeper understanding. We assume that the reader has some familiarity with calculus, linear algebra, elementary probability, and basic statistics.

In Chapter 1, we give a brief introduction to the basic ideas of algebraic topology, including discussion of algebraic background (linear algebra and abstract algebra), basic point-set topology, simplicial complexes, and the construction of homology groups. In Chapter 2, we give an overview of topological data analysis, focused on the theory surrounding persistent homology. We review the machinery for understanding topological invariants of data sets in terms of associated simplicial complexes, explain persistent homology and the basic structural theorems, and describe the Mapper algorithm. In Chapter 3, we describe the emerging and active area of research integrating topological data analysis with the methods of statistics; this is a necessity for the use of these tools to analyze scientific data and perform inference. In Chapter 4, we give a brief overview of the area of manifold learning, which is closely related to topological data analysis, and review mathematical models of spaces of phylogenetic trees.

In the second part of the book, we explore some biological applications. In Chapter 5, we study the topology of point clouds in genomic space using persistent homology and the geometry of phylogenetic spaces. Specific examples include viruses (influenza and HIV), bacteria, and humans. Chapter 6 provides a concise introduction to cancer genomics; among the applications, we use topological data analysis to study the evolution of tumors in collections of patients, to describe the stratification of patients, and to capture the association between genomic data and sensitivity to diverse therapeutic agents. Next, in Chapter 7, we turn to a new type of data that is particularly well suited to TDA tools: expression profiles of large collections of single cells. In Chapter 8 we study the three dimensional structure of DNA using persistent homology, with examples from bacteria and human cells. Finally in Chapter 9 we use a mapping of time-series data into finite metric spaces to extract periodic features. Each of these chapters contains background information on the relevant biological problem and can be read independently.