9

Topological Data Analysis beyond Genomics

In this last chapter we will briefly introduce several recent interesting applications of TDA to diverse biological problems beyond the genetics and genomics work we have focused on in this book. In the first part of this chapter we will explain how TDA can be used to study ordered data, referred to here as series data. Series data is frequently found in many biological applications; for instance, when studying the evolution of a biological organism or population, where data is ordered in time, or when looking at genomic data along a chromosome, where data is ordered by chromosomal location. Examples of time series data with periodic patterns can be found in the cell cycle, or the phenotypic changes in immune genes following infection and recovery [501].

Next, we will discuss TDA techniques for studying graphs, or networks. Networks are standard representations of complex biological systems with different components interacting. For example:

1. The set of interactions between different proteins within an organism is traditionally represented by a graph where nodes represent proteins and edges physical interaction.
2. Transcriptional networks are captured by graphs, where nodes represent genes or transcripts and edges represent how the expression of one relates to the expression of the other.
3. In neuroscience, neurons and their interaction are usually encoded by a graph. The central problem in neuroscience focuses on how the neuronal system captures information about the world; a key question is to how to study this problem using the structure of the interaction graph.

As graphs are pervasive representations of biological data and the simplicial complexes that arise in TDA are generalizations of networks, it is perhaps unsurprising that there are interesting TDA approaches to extract and quantify global properties of biological networks.

Next, we describe some natural sets of applications for TDA in medical imaging. For example, magnetic resonance imaging (MRI) is a non-invasive technology that is used to test for many diseases and to obtain data on the real-time activity of living organisms. The output is a function from the three physical dimensions of space (and time if dynamic information is being captured) to the real numbers. Filtrations of this function can be studied using TDA methods and one hopes to relate these topological features to biologically or clinically interpretable characteristics.

Finally, we briefly mention some recent applications of TDA in the context of infectious diseases: first, models of networks of infectious disease spread in a population; second, how organisms respond to infectious diseases.

Our goal in this chapter is to provide a very brief overview of examples of TDA techniques applied to a variety of biological problems beyond the main scope of the book. Our treatment is of necessity superficial, and in particular by no means should be viewed as comprehensive. (We apologize now for work which is omitted; our choices here are not intended to reflect a judgement about the most interesting work.)

9.1 Topological Study of Series Analysis

Time series analysis is an old discipline aiming at extracting patterns and summaries from data arising from weather measurements, financial markets, signal processing, and many other systems. Biological processes are not an exception. In many biological problems data is naturally ordered along a well-defined physical or biological dimension. For example, the position of genes along a chromosome specifies an ordering. Another set of examples come from the time course of a biological process (Figure 9.1).

The first applications we will describe here are time series analysis of expression data. There are a large variety of biological systems that display interesting time dependent expression profiles. For instance, periodicity is observed in circadian regulation, the cell cycle, and the life cycle of malaria [6], among many other examples (Figure 9.2). Genes are regulated according to different temporally orchestrated transcriptional programs, and discovering information about this time dependence can inform theories of how these programs are organized. Several techniques have been used to study time series expression and recently some implementations using ideas from topological data anlysis have appeared. Cohen-Steiner and colleagues [118, 140] proposed a measure of similarity of expression profiles of different genes based on comparing the persistence diagrams arising from level-set persistence (recall Example 2.3.4) applied to a function from time to expression

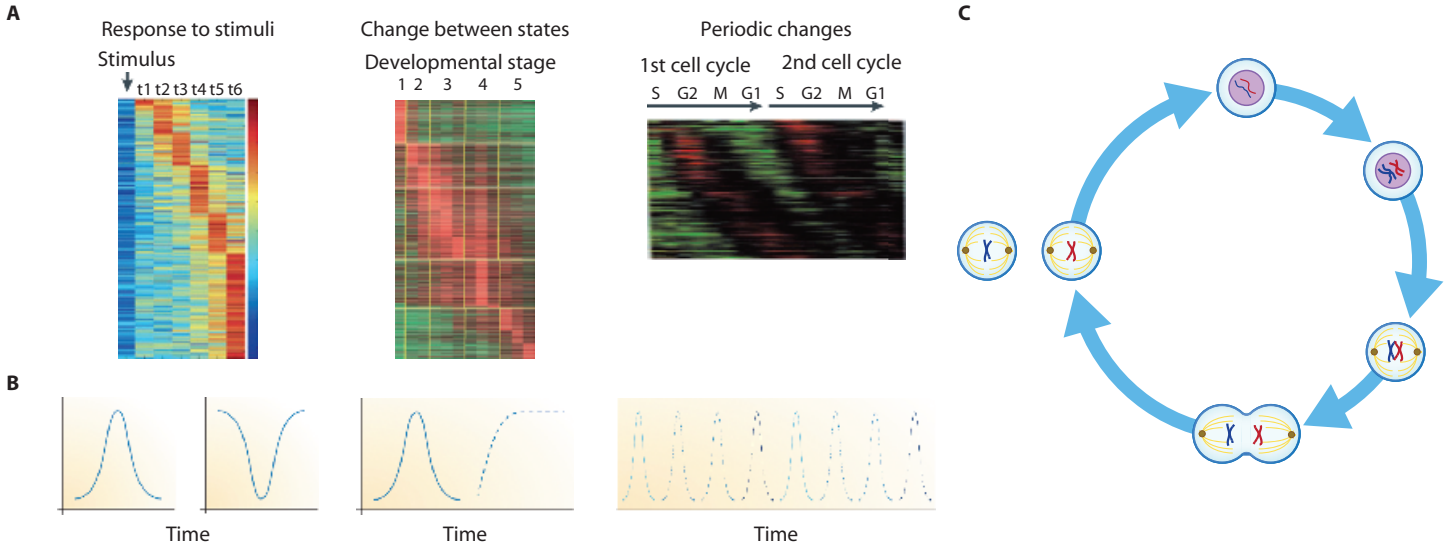

Figure 9.1 Changes in transcription are observed in multiple biological processes. (A) Time series data is ubiquitous in biological processes: for example, response to transitory external stimuli, changes between two different states (in development, for instance), or cyclic changes as observed in the cell cycle or circadian rhythm. There are many interesting biologically interpretable patterns of potentially infinite types. (B) Qualitative phenomena include pulses (a single spike associated to stimuli), sustained changes, periodic changes, among many other examples. (C) An example of a fundamental cyclic process: the cell cycle, where DNA replicates and a mother cell divides into two daughter cells. Source: [35]. Reprinted by permission from Springer Nature: Springer Nature, *Nature Reviews Genetics*, Studying and modelling dynamic biological processes using time-series gene expression data, Ziv Bar-Joseph, Anthony Gitter, and Itamar Simon, 13.8 (2012): 552–564. © 2012.

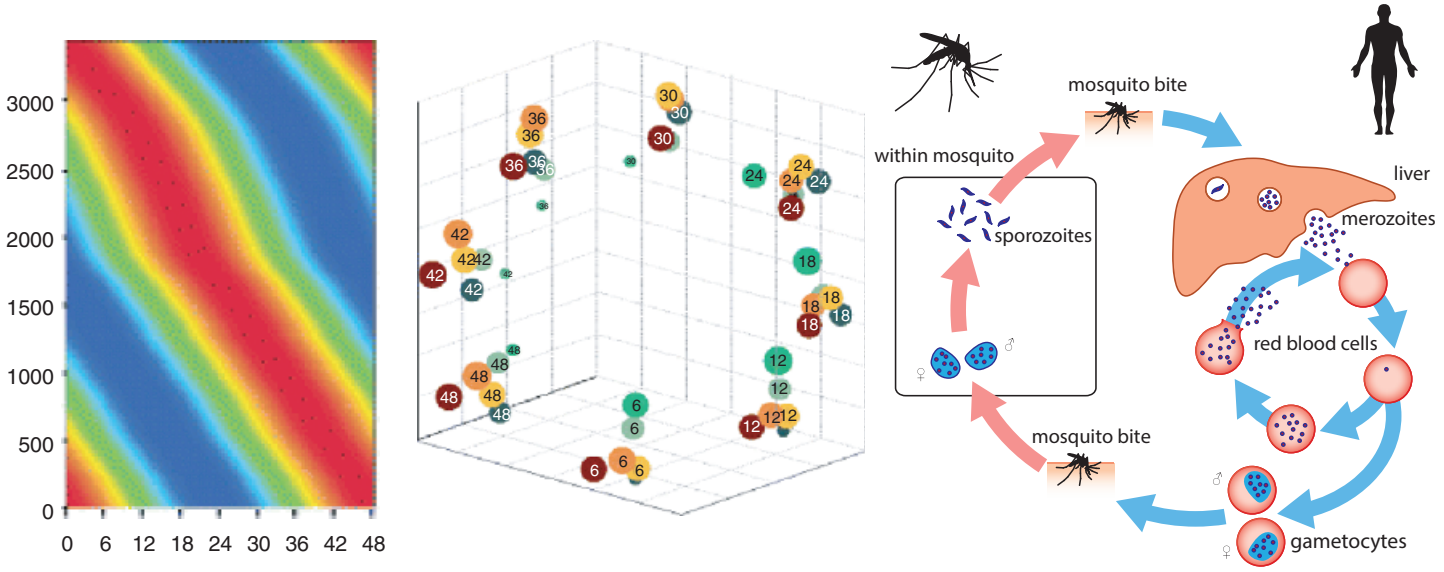

Figure 9.2 Expression of different genes in the malaria parasite life cycle. Left: Genes can be ordered by the time of expression. Right: Using PCA (see Section 4.2 for a brief overview) one can observe a cycle in gene expression reflecting the parasite (*Plasmodium falciparum*) life cycle. Time series data (expression of *P. falciparum* genes at different time points) was analyzed using the fast Fourier transform (FFT) and enrichment of different cell processes. The higher PCA components replicate the cycle of *P. falciparum* replication. Source: [6].

levels. Perea and Harer [404] proposed a method based on a common strategy in time series analysis, applying a sliding window. As we explain below, they regard the sliding window as a map from the time series data to point cloud data, and then explain how to use topological properties of this point cloud to study the periodicity of the original time series data. (There has also been interesting recent work by Khasawneh and Munch [291] on stochastic delay differential equations.)

Finally, we turn to an application to data that uses the natural ordering in genomics coming from the position along a chromosome. As we saw in Chapter 6, chromosomal aberrations (deletions, amplifications and translocations) are very common events in most tumors. The number of copies of particular chromosomal regions is a function of the ordering of genes. Arsuaga and colleagues have also proposed a sliding window approach to study copy number aberrations in cancer [19, 21]. The sliding window, moving across the chromosome, provides a map from the copy number data to point cloud data; hierarchical cluster structure in this point cloud data, as measured by the zeroth Betti number, reflects changes in copy number structure. This approach can be used to identify copy number changes in tumors and to compare the profile of these changes across different tumors.

9.1.1 Time Series Analysis of Gene Expression Data

Biological processes are dynamic, changing at different time scales. For instance, we could be interested in tracking the expression of a particular gene that is involved in the cell cycle. As we saw in Chapter 6, the cell cycle is one of the fundamental processes altered in cancers, and looking at how proteins differentially altered the cell cycle in cancer cells could provide therapeutic opportunities. Another example is the circadian rhythm, which regulates fundamental biological processes in a daily cycle. These biological processes can be studied by measurements $x(t_i)$ taken at different times $\{t_i\}$; this is popularly called a time series. There is an extensive literature on general methods for time series analysis [102, 226], but of particular interest in this chapter is the identification of periodic signals in gene expression data.

Time series expression data presents some distinctive features that make it different from finance or weather time series: it is usually collected for a few cycles, the sampling can be sparse and uneven, there is not usually a characteristic shape (such as a sinusoidal curve), there is significant biological variability and, last but not least, there is a potentially significant amount of experimental noise. All these features make time sampling of genomic data particularly interesting and challenging.

Since the first high-throughput expression experiments at the beginning of the century, there has been a plethora of methods applied to extract signals from time

series expression data. These methods usually rely on standard techniques used and developed in other fields, for example Lomb-Scargle periodograms developed for astrophysics [202, 333]. We will briefly summarize some of the most common techniques used for time series analysis to study biological genomic/transcriptomic data.

- **Spectral methods**. Spectral methods express signals in terms of the frequency domain (e.g., via Fourier analysis). A basic and widely used method is direct application of the fast Fourier transform (FFT) algorithm, which transforms discrete data into Fourier components. A particularly useful technique when working with uniformly spaced time-sampled data is to approximate the spectrum by the periodogram:

$$s(\omega) = \frac{\Delta t}{N}\left|\sum_{n=0}^{N-1} x(t_n)e^{2\pi i n\omega/N}\right|^2$$

where Δt is the time interval between two observations and N is the total number of observations. Periodic signals can be identified as sharp peaks in the periodogram.

Fourier analysis has been widely applied to study cell cycle genes from expression data (e.g., [537]). Fourier analysis in connection with permutation tests was implemented in [136] and applied to yeast cell cycle data and other species [272].

However, the FFT is suboptimal for sparse and non-uniformly sampled data. The Lomb-Scargle periodogram [333, 448] is a Fourier type of analysis able to infer spectral properties from sparse and irregular sampling at times t_k [202]. For a fixed frequency ω, a time delay τ is defined as a solution to the equations

$$\tan 2\omega\tau = \frac{\left(\sum_k \sin 2\omega t_k\right)}{\left(\sum_k \cos 2\omega t_k\right)}.$$

Then the periodogram at the frequency ω is equal to:

$$s(\omega) = \frac{1}{2}\left(\frac{\left|\sum_{k=0}^{N-1} x_n \cos\omega(t_k-\tau)\right|^2}{\sum_{k=0}^{N-1}(\cos\omega(t_k-\tau))^2} + \frac{\left|\sum_{k=0}^{N-1} x_n \sin\omega(t_k-\tau)\right|^2}{\sum_{k=0}^{N-1}(\sin\omega(t_k-\tau))^2}\right).$$

The Lomb-Scargle periodogram has been used in several biological applications with incomplete data or time series sampled at different time points [422, 441], including the cycle of malaria [202], circadian rhythms in plants [271], and phenotypic behavior in animals [286], among many others. For instance, in reference [202] it was used to study the expression of *Plasmodium falciparum* (the agent causing malaria) genes, during infection. The Lomb-Scargle periodogram

analysis showed higher sensitivity than the Fourier transform in the identification of periodic signals, mostly in day and two day cycles.

- **Wavelets**. The Fourier transform decomposes the temporal data as a sum of orthogonal sinusoidal representations; this is typically used for identifying periodic patterns in data. Wavelets provide an alternative decomposition in terms of an orthogonal basis of multiresolution functions (wavelets) $\psi_{j,k}(t)$ localized around a time-frequency region parametrized by the k and j indices; the basis elements are scaled and shifted versions of a generating "mother wavelet." The wavelet basis allows us to express any function as follows:

$$x(t) = \sum_{j,k} c_{j,k} \psi_{j,k}(t)$$

where $c_{j,k}$ are the coefficients. (There are many different choices of wavelet bases, including Haar and Daubechies wavelets.)

Wavelets have been used to study clusters in gene expression along the genome [509], regulatory networks from time-varying expression data [189, 295, 475], and functional MRI data [446].

- **Reference curve comparison**. Fourier analysis decomposes data into sinusoidal signals, but the data that we are interested in could have different shapes. For instance, we can be interested in identifying narrow peaks indicating the expression or activity of a particular gene in a narrow time window. If we have a particular time dependence in mind, regression analyses could provide useful insight. Partial least squares regression (PLS) has been used to identify genes with periodic expression along the cell cycle in *Saccharomyces cerevisiae* [275]. In that work, the authors were interested in the identification of periodic signals with a common period (cell cycle) but where different genes obtained the highest expression at different times of the cell cycle. That was modeled by a function $A \sin(\omega t + \phi)$, where $\omega = 2\pi/T$ is the frequency associated to the cell cycle. Every gene received an amplitude and a phase, representing the variability along the cell cycle and the cell cycle phase. The same procedure can be used for a non-sinusoidal family of curves, such as the ones shown in panel B of Figure 9.1, using PLS to find the parameters corresponding to each family member.

The Jonckheere-Terpstra trend test [278] is a non-parametric statistical test for comparing two alternative hypotheses regarding the medians of populations; the null hypothesis is that the medians are the same, and the alternative hypothesis is that the ordered populations have increasing medians. This is closely related to the Kendall τ statistic for analyzing rank correlation. JTK CYCLE is an algorithm based on these statistics that compares data to a set of hypothesized user-defined group orderings [258]. The algorithm has been extensively applied to study different aspects in different systems of circadian rhythms including expression profiles, chromatin changes, metabolic changes and changes in

microbiota among many others [132, 258, 303, 496]. An advantage of these methods is that since they are rank based, they are robust invariants and hence resistant to corruption by outliers (recall the discussion from Chapter 3).

- **Stochastic processes and correlograms**. A stochastic process is a set of ordered random variables $x(t_i)$. There is an extensive literature on the study of time series analysis using stochastic processes [102]. Here, we will briefly mention autoregressive models (AR) of order p as a common type of stochastic model used in time series analysis. The main idea behind these models is that the value of the observation at time t_i depends on a linear combination of the previous p-values $\{x_{i-k}\}$ for $k = 1, \ldots, p$, in other words:

$$x(t_i) = \sum_{k=1}^{p} c_k x_{i-k} + \epsilon_{t_i}$$

where c_k are some real coefficients and ϵ_{t_i} is an error term, typically assumed to be Gaussian distributed. These are a generalization of Markov processes; in fact, AR(1) models are precisely Markov processes.

A generalization of AR is given by the moving average process (MA), where the value of the observation depends on a linear combination of q independent random processes ϵ_i with zero mean and equal (finite) variance:

$$x(t_i) = \sum_{k=1}^{q} c_q \epsilon_{i-q}$$

The most general models contain a sum of p terms from an AR model and q terms from an MA model; these are usually called (p, q) ARMA models.

One common assumption in stochastic processes is that the observations $x(t_i)$ are derived (possibly after subtraction of global trends) from a *stationary* stochastic process. A stochastic process is stationary if the joint distribution of every set of variables $x(t_a)$ is the same as $x(t_a + \tau)$ for all τ. In other words, the process does not present any systematic change in mean, variance and higher moments at different time points. Stationarity is a very strong assumption, and it is usually only applicable after all trends (changes in mean, variances, periodic components) are removed from the original data.

A useful function for studying stationary processes is the autocorrelation $\gamma(\tau)$, defined as the covariance between $x(t_i)$ and $x(t_i+\tau)$ divided by the value at $\tau = 0$. Using the data of a stationary process one can define [102]

$$r(k) = \frac{\sum_{i=1}^{N-k} (x_i - \bar{x})(x_{i+k} - \bar{x})}{\sum_{i=1}^{N} (x_i - \bar{x})^2}$$

where N is the total number of observations. The function $r(k)$ is called the correlogram and carries interesting information on the coefficients and memory

of stationary stochastic processes. These models have been proposed for gene clustering from time series of expression data [188, 418, 525].

- **Compressibility**. Most patterns in biological data do not have a predefined form; the use of reference curve comparison could preclude the identification of some relevant biological signal. One idea for capturing general regularity is to study the compressibility of the data. One can define the algorithmic complexity of a string of characters as the size of the shortest program that outputs the string [304, 473]. These ideas were applied to yeast cell cycle expression data in [10]. First, the data were transformed to their rank value, for instance, 2.5, 2.7, −1.2, 23 will be transformed to 2, 3, 1, and 4. This corresponds to a permutation of 1, 2, 3, and 4. Then one looks at a function f from the permutations to the real numbers. For example, such a function could be the length of the longest increasing or decreasing sequence (2 in our case), the number of local maxima (2 in our example), the sum of the absolute values of the difference between consecutive numbers ($|2-3|+|3-1|+|1-4|=6$), etc. (Notice that some permutations will be assigned the same value.) One way of describing the permutation p of interest (e.g. 2, 3, 1, and 4) is to count the number of permutations with the same image under f; denote this quantity by M_f. A bound on the compressibility can be obtained by $k(f) = \log M_t - \log N - \log M_f$, where M_t is the total number of permutations and N is the number of different values the function f can take. The main idea of the method of [10] is that simple functions can be used to identify interesting patterns as corresponding to highly compressible permutations. Note that these patterns are not periodic.
- **Biologically based time dependent models**. In some cases, there are actual models intended to describe the evolution of the biological system. For example, transcription is a classical problem where different models have been proposed that relate the activity of some genes (transcription factors) in the regulation of other genes. Such models are usually represented in the form of a network (e.g., Boolean or Bayesian networks or sets of first order differential equations) [34, 296].

All these methods have tradeoffs [137]. If we are interested in looking for periodic signals using a large collection of longitudinal data, we might be tempted to try Fourier approaches. On the other hand, if we are studying a narrow localized signal, we could try some of the simple wavelet methods. Stochastic methods are useful if we have reason to believe that the hypotheses of the main methods (e.g., AR, MA, ARMA, etc.) hold in data; for instance, when the noise can be modeled with a known distribution and there is a memory of a few previous values. In other cases, we might have a suspicion of what kind of signal we should expect and we can try to fit the expected curve directly to the data.

Of course, in practice, especially when performing exploratory data analysis, we do not expect to have a clear idea of what kind of information we are looking for. The topological approaches that we will describe in the following section provide a more general framework for identification of periodic patterns. At the moment it is an open question which methods will be most informative when working with the time series data arising in genomics. As transcriptomic data becomes more reliable and abundant, we expect to have the opportunity to evaluate the performance of these algorithms.

9.1.2 Time Series Analysis Using Topological Data Analysis

A simple mathematical model of the expression values of a particular gene i evolving in time is simply a function $e_i \colon X \to \mathbb{R}$, where X could be an interval $[a, b]$ (when considering a fixed time interval) or S^1 when looking at periodic systems. Of course, in practice we expect to have access to the values of these functions at a finite collection of times (i.e., points of the domain).

A first natural question is how to compare the expression patterns of two different genes i and j; in this model, we are comparing the functions $e_i(t)$ and $e_j(t)$. One way to answer this question is to consider the sublevel set filtration of $e_i(t)$ and $e_j(t)$ (recall Example 2.3.4); we consider the filtration of spaces induced by considering the collection of sublevel sets $f^{-1}((-\infty, a])$, which are equipped with evident inclusions

$$f^{-1}((-\infty, a]) \to f^{-1}((-\infty, a'])$$

for $a < a'$.

The two functions $e_i(t)$ and $e_j(t)$ can then be compared by measuring the bottleneck or Wasserstein distances between the persistence diagrams arising from this filtration. The stability theorems for persistent homology in this context (recall Theorem 2.4.12) now imply that these measures are fairly robust in the face of sampling variation or noise in the data (Figure 9.3).

Sublevel set persistence and bottleneck distances were used in [118, 140] to study clustering of genes by expression level over different developmental stages in microarray data from the arabidopsis plant. Specifically, the data is structured as vectors of expression levels for each gene, with entries corresponding to developmental stages. In the same papers [118, 140], the authors used topological methods for identification of periodic signals using expression values of 7500 genes across 17 time points within a single period of the formation of somites in mouse embryo. In combination with a variety of other methods (e.g., Lomb-Scargle periodogram

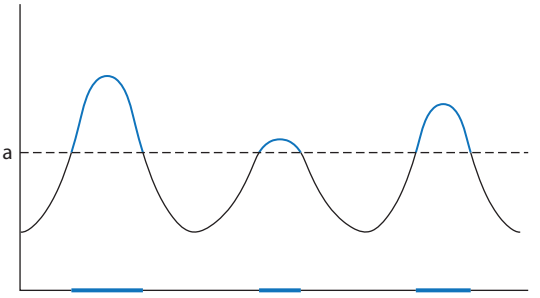

Figure 9.3 Persistence homology can be applied to filtrations from different sublevel sets induced by a function.

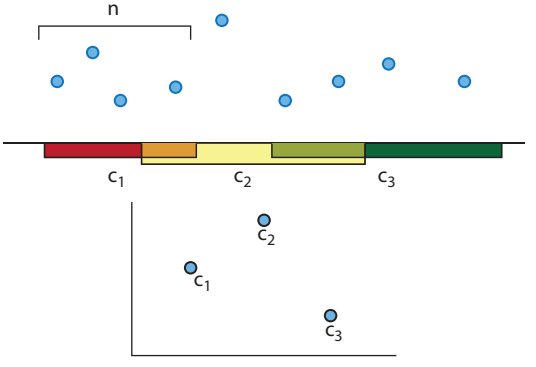

Figure 9.4 The sliding window approach is a very common strategy used in many genomic analysis applications. Data corresponding to a window of constant size n can be represented as a point in n dimensions. Sliding the window, one can generate a point cloud data representing the series. TDA techniques can be then applied to the cloud data to learn properties of the series.

and the cyclohedron method), this study found a new cyclic gene that regulates the segmentation clock. Comparative analysis found the topological methods to be competitive with other methods, although not obviously superior to the best alternatives.

9.1.3 Topological Data Analysis of Sliding Windows

A different strategy of studying time series data is to study the point cloud data generated by sliding windows (Figure 9.4). An illuminating analysis and development of this method for periodicity detection was carried out by Perea and Harer [404].

The basic idea is simple. Suppose we have a series of points $\{x_1, x_2, \ldots, x_n\} \subset \mathbb{R}$, where the subscript indicates the time label of the point. We think of these as the

image of a function $f\colon \mathbb{R} \to \mathbb{R}$ on a collection of values $\{t_1, \ldots, t_n\}$. We define a *window* of size w starting at interval i as the collection of points

$$\{f(t_i), f(t_{i+1}), \ldots, f(t_{i+w-1})\} = \{x_i, x_{i+1}, \ldots x_{i+w-1}\}.$$

An idealized case is where $t_i = t_1 + (i-1)d$, for some shift value d.

As an abstraction, imagine that we simply have a function $f\colon \mathbb{R} \to \mathbb{R}$, and when fixing a window starting at value x we parametrize by M and τ and instead consider the collection of points

$$\{f(x), f(x+\tau), f(x+2\tau), \ldots, f(x+M\tau)\}.$$

Fixing M and τ, we can regard the window as specifying a curve

$$W(f)_{\tau,M}\colon \mathbb{R} \to \mathbb{R}^M.$$

To understand what this looks like for a periodic signal, it is interesting to focus on $f(\theta) = \cos(L\theta)$ for some choice of period $L \in \mathbb{N}$. An easy analysis shows that the resulting closed curve traces out an ellipse and that the length of the minor axis of the ellipse is maximized when the window size is close to the period!

This suggests the approach of computing the length of the longest barcode in the persistence diagram for H_1 as an estimate of the periodicity; we can in fact recover the period exactly in this case. More generally, any suitably bounded function can be expressed via the Fourier transform as a linear combination of periodic functions; the stability theorem for persistent homology can then be used to show that in the case where this signal is periodic, the longest barcode still recovers useful periodicity information.

Based on this analysis, Perea and Harer propose the algorithm SW1PerS and demonstrate applications to finding periodicity in gene expression from yeast metabolic and cell cycles [405]. In simple tests, the algorithm compares well to existing tests for periodicity (notably Lomb-Scargle). In particular, SW1PerS has extremely good performance in the face of noise, performing better than Lomb-Scargle in high noise regimes on signals where the magnitude decays over time and where there are separated peaks.

9.1.4 Identification of Copy Number Alterations

Time series are not the only interesting ordered biological data sets. Chromosomes present a natural one-dimensional ordering of genes. In cancer, chromosomal regions are often amplified, deleted, and translocated. Regions that are recurrently amplified could contain oncogenes, regions that are deleted could contain tumor suppressors, and regions with translocations could give rise to gene fusions. For instance, many tumors contain deletions in the 9p21.3 region containing a known

tumor suppressor gene *CDKN2A*, or the region 17p13.1 containing the *TP53* gene. A common approach for the identification of genes that could be implicated in cancer is to assess recurrence of alterations across many different patients. Most methods that have been proposed using copy number alterations in cross-sectional samples propose a measure of recurrence and a statistic associated to it [52, 508].

Arsuaga and colleagues have proposed a method for the analysis of copy number data using persistent homology [19, 142]. The idea is based on a sliding window approach similar to the method used for studying periodic signals in time series [404], but instead of using time as the ordering dimension they use the chromosomal position. A sliding window defines a map from the copy number data to a w-dimensional space. In this case, the authors chose a three-dimensional window, so the data can be viewed as point cloud data in three dimensions. If there are no copy number alterations, one should expect that the data should fluctuate around the expected number (two in the case of autosomes), so this will correspond to fluctuations near the point $(2, 2, 2)$ in the three-dimensional point cloud. However, a deletion will change this number to one (heterozygous) or zero (homozygous), changing the concentration point for the point cloud to $(1, 1, 1)$ or $(0, 0, 0)$. In the same fashion, amplifications could be identified as changes in the point cloud data to concentrate around other diagonal values.

The authors use the zero dimensional persistent homology (i.e., the dendrogram representing single-linkage clustering) for the identification of regions containing copy number alterations. To derive a statistical test, the resulting PH_0 barcode was compared to one generated by a non-tumor control. One should expect that controls will become a single cluster at small values of the filtration value whereas the tumor samples containing copy number alterations will cluster in several groups at low filtration value and become a single connected component at larger filtration value. To assess the statistical significance, Arsuaga and colleagues propose a statistic $s = \Sigma_\epsilon (t_\epsilon - c_\epsilon)^2$, where t_ϵ and c_ϵ denote the average number of connected components in the test and control data sets, respectively. An associated null hypothesis (and p-value) was generated by random permutations of the data. When applied to breast tumors from an independent study [249], the authors were able to recapitulate known recurrent alterations and to identify some unreported alterations.

The same approach could be used for the identification of regions of differential gene expression. Using chromosomal position as ordering dimension, and expression of genes as a function, Arsuaga and colleagues [21] generated point cloud data using a sliding window approach. Notice that expression of a gene is correlated to copy number information, i.e., highly amplified genes tend to be more expressed, and deleted genes less expressed. The number of clusters as measured by

H_0 is associated to consistent changes in expression profiles across a chromosomal region. Applying this approach to 251 breast cancer expression profiles from [353] identified specific clustering profiles associated to different expression subtypes.

9.2 Topological Data Analysis in Networks and Neuroscience

Networks have become a common representation of many biological systems, representing different scales of knowledge. For instance, protein-protein interactions are summarized as networks where we can represent proteins in a living system as nodes and their interactions as edges. Neurons in the brain and their interactions provide another example of a biological system where some properties could be loosely captured by networks. The architecture of the brain as captured by the interconnections between different regions provides another example. Researchers are currently exploring the use of topological techniques to characterize the molecular, neuronal, and architectural properties of the brain [98, 130, 201, 407, 424, 463]. The hope is that topological techniques will provide ways to summarize properties of complex networks that generalize and extend the standard invariants based on global statistical properties of local information (e.g., degree distribution, measures of centrality of a vertex, or number of components).

9.2.1 Cellular Scales: Neuronal Activity

One of the central problems in neuroscience is how the ensemble of neurons can efficiently capture and faithfully represent information about the world. Neurons in physical proximity can exchange information across synaptic gaps, reflected in their neuronal activity. The relationship between neuronal connectivity and activity is highly nonlinear. Linear techniques do not suffice to correctly capture its structure. The visual cortex is a classical system in which to study the coding of external physical stimuli into neurons. Singh and colleagues [463] used persistent homology to study the population activity in the primary visual cortex. The invariants derived from persistent homology in natural image stimulation were similar to a spontaneously active cortex. Giusti and colleagues [201] proposed a method based on TDA to extract nonlinear but monotonic relationships. The hippocampus has been found to encode information about the physical environment through pyramidal neurons. Different neurons in the hippocampus respond selectively to different physical locations [390]. In [201], the method is shown to be able to recover geometric information encoded in neuronal correlations (without using the external stimuli); that is, they can recover place cell activity without hypotheses about the stimuli or the receptive fields of the cells.

Reimann et al. [424] explored the relation between neuronal architecture and information processing by constructing directed graphs capturing the direction of synaptic transmission. In particular, they summarized data as a directed graph with nodes representing individual neurons and directed edges representing pre- to postsynaptic neuronal connections. The response to external stimuli can be modeled as time series data in the directed graph. Different aspects of the structure of these graphs can be quantified by identifying different objects at different scales, from local (as indicated by the presence of cliques of neurons) to global (as indicated by the existence of larger topological structures). (See Figure 9.5.) Applying TDA techniques to computational reconstructions of neocortical circuits in the brain of a rat, the authors were able to quantify the presence of these different structures, including large numbers of high-dimensional cliques and "holes." This quantification of structural properties of neuronal networks hopefully provides a first step for understanding the association between brain architecture and function.

9.2.2 Mesoscopic Scales: Brain Functional Networks

Cognitive processes usually involve the coordinated activity of different areas of the brain. The relation between the activity of brain regions can be represented by networks, and statistical properties of these networks can provide information on the functional architecture. Functional imaging can provide information on the activity of the brain at mesoscopic scales of thousands or millions of cells. Petri et al. [407] proposed to use TDA techniques to study the statistical properties of homological cycles in these networks. To test these ideas, they compared the resting state of 15 healthy volunteers receiving placebo or psilocybin, a psychoactive drug. A significant difference was observed in the homological features between the two groups, suggesting that these rough descriptors capture relevant structural properties of brain architecture. The brain architecture of structural connectomes was also explored using topological techniques by Sizemore et al. [464]; in particular, they sought to identify densely connected groups of active regions. Further, they proposed that these cliques were related to local fast processing. Experimentally, they verified that these regions were consistent across a group of eight individuals. Cassidy et al. [98] has applied TDA to compare the activity of the brain using functional MRI (fMRI) in a variety of conditions. This approach improves over standard correlation comparison methods, in the sense that when applied to real data it produces much more sensible functional connectivity predictions. It involves a method to analyze network architecture using persistent homology, accounting for potential artifacts due to spurious spatial and temporal correlations.

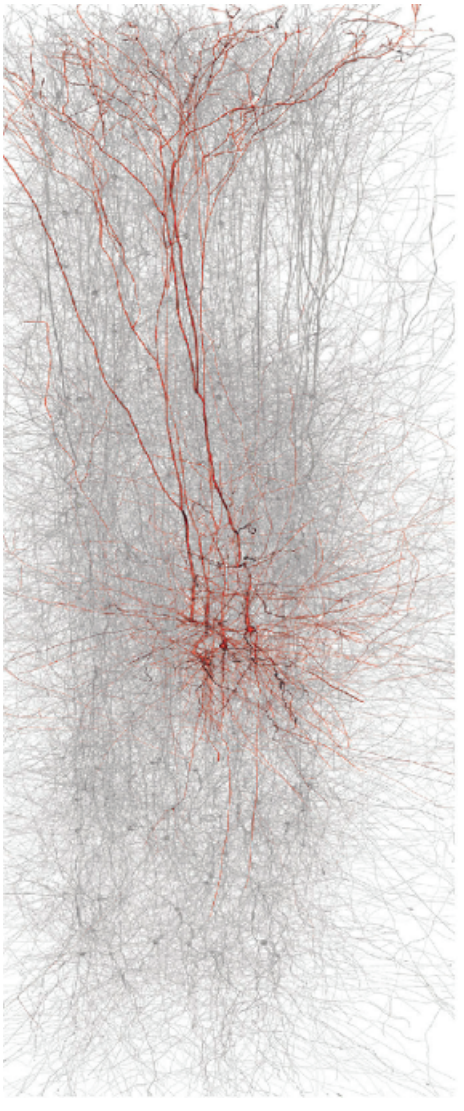

Figure 9.5 Representation from a slice of in silico reconstructed neuronal tissue. In red, a clique formed by five pyramidal cells. Source: [424].

9.3 Topological Approaches to Biomedical Imaging

Imaging is one of the main non-invasive modalities for diagnosing and evaluating the progression of many diseases, including cancers. Solid tumors appear as masses in various imaging technologies, notably including MRI. Tumors have a shape and volume; these can be modeled as the topological and geometric properties of a three-dimensional object. Rough metrics on these masses (e.g., changes in

volume) are used as a standard for prognosis and to evaluate therapeutic efficacy. However, it has been found that other geometric invariants can provide interesting clinical information. For instance, glioblastomas, the most common type of brain tumors in adults, come in two types: one single mass or multiple masses (multifocal/multicentric glioblastomas). That is, a glioblastoma is classified by whether it has multiple path components. In [320], it was shown that these two types are associated to different genetics: multifocal/multicentric tumors are strongly enriched in point mutations in *PIK3CA*, a major oncogene, and are genetically highly heterogeneous with different lesions associated to different masses in the tumor (Figure 9.6). This observation has important clinical implications, as multifocal/multicentric glioblastomas have a worse prognosis and drug responses to different masses are extremely heterogeneous.

The observation that simple geometric and topological properties of images (volume or number of path-connected components) can inform prognosis and drug responses prompts the question of how image data relates to genetic and clinical data. Ideally, one would like to systematically explore the map between genetic and phenotypic data (as expressed in the image and in other clinical sources).

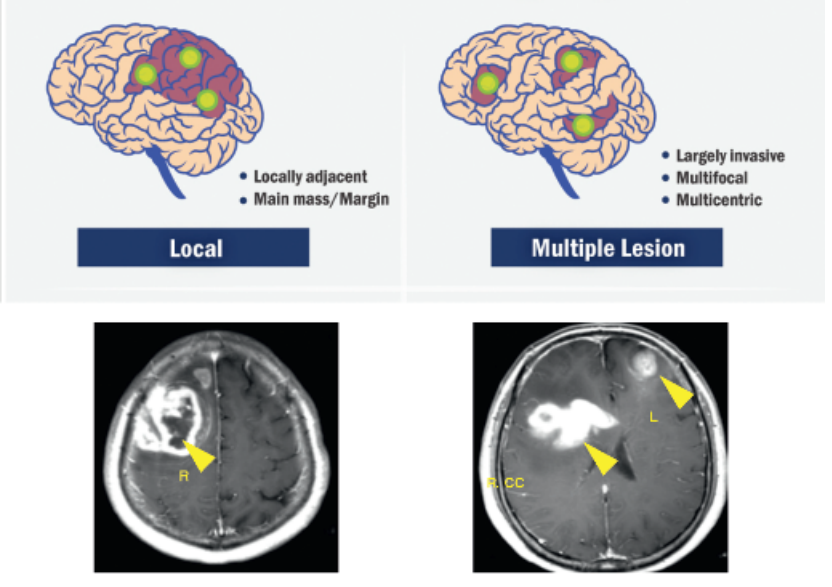

Figure 9.6 Glioblastomas can appear in a single mass (left), or several masses (multifocal/multicentric glioblastomas) (right). This simple topological difference is associated to specific mutations (*PIK3*) and worse prognosis. Source: [320]. Reprinted with permission of Springer-Nature: Lee, Jin-Ku, et al. "Spatiotemporal genomic architecture informs precision oncology in glioblastoma." Nature Genetics 49.4 (2017): 594–599.

Crawford et al. [128] proposed to use the smooth Euler characteristic transform (recall Section 3.8) to decompose tumor image data into a set of topological features amenable to machine learning. The procedure starts by segmenting the tumor image (i.e., identifying from the image data the tumor and reconstructing a three-dimensional manifold [112]), then sectioning it using the smooth Euler characteristic transform to extract a function of different directions that can be used for subsequent functional machine learning analysis. Crawford et al. showed that topological information provides complementary information to other types of genomic, transcriptomic, and volumetric data to predict overall survival in glioblastomas. This work suggests an interesting approach of combining "omic" data with imaging to better characterize the mechanisms of initiation and progression of tumor growth; topological analysis appears naturally as a way of quantifying imaging features.

Another interesting complex network studied using TDA is the blood vessel system. In [489], Szymczak and colleagues proposed reconstructing the vascular trees from three-dimensional images using persistent homology. The method was applied to reconstruct coronary trees from computed tomography (CT) scan data of the heart.

9.4 Spreading of Infectious Diseases

Networks have also been used to capture the spread of infectious agents, where nodes represent infected individuals and direct infection is represented by edges. Understanding the structure of these networks is one of the main objects of study in epidemiology; such analysis provides information about the main routes of transmission (aerosol, food, through other species, etc.), spread via local contact versus long range contact (e.g., airline influence), how effective is the transmission of infection, potential sensitive and resistant populations, and the efficacy of potential ways of curtailing the spread. Taylor and colleagues [491] used TDA techniques to study the mathematical structure of these graphs, their intrinsic dimensionality, and their topological and geometric properties; they connected these invariants to epidemiologically significant quantities.

Another application is the study of immune responses to infectious agents [501]. It is well understood that different hosts will have very different responses when exposed to the same infectious agents: some are resilient, others present mild symptoms, and others could die. Torres et al. propose a phenotypic space, the "disease space," that captures the potential states of an infected host (Figure 9.7). Measuring physiological data (weight, temperature, different cell counts, among others) at different states of infection, one can trace the trajectories of

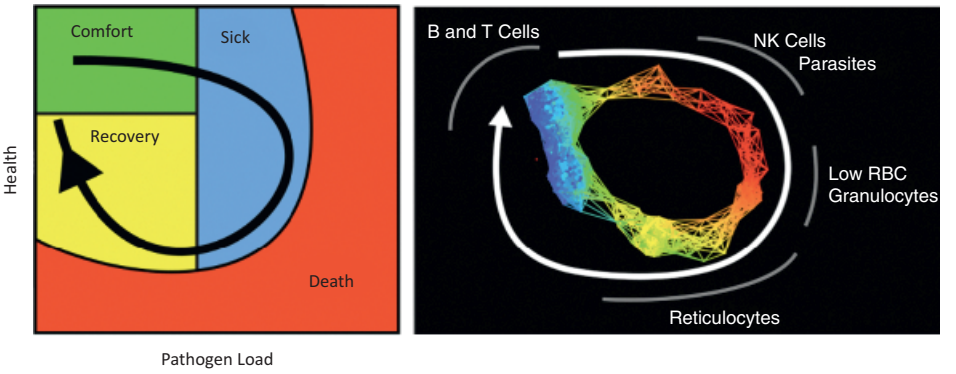

Figure 9.7 Left: Conjectured "disease space" capturing the potential phenotypic states of a host infected with a pathogen. The normal state of an uninfected individual is on the left. When it is exposed it is thrown out of this state. As the host recovers, it goes back to the healthy state through a trajectory that does not track back the previous states. Right: Mapper applied to physiological data from mice exposed to *Plasmodium chabaudi*. Source: [501].

different hosts after exposure. It was observed that resilient hosts do not significantly change states, while less resilient hosts are associated to large "loops" in the disease space. These hypotheses were evaluated in mice exposed to *Plasmodium chabaudi* (a murine analogue of the cause of human malaria) and in malaria patients.

9.5 Summary

There are many exciting directions in the application of TDA methods to biological data beyond genomics, and we expect more to be discovered in the coming years. This work is clearly in its infancy, but already there have been interesting and suggestive results.

- There is a great deal of biological data which comes as an ordered sequence; time series data is a notable example, but not the only one.
- Topological data analysis methods provide a way to detect periodicity in ordered signals, via a sliding window approach, that is competitive with the best standard methods (and has different properties).
- TDA methods for analyzing graph structures have been profitably applied to problems in neuroscience (studying neuronal connectivity and its correlation with neural activity), epidemiology, and analysis of coronary artery structure.

9.6 Suggestions for Further Reading

- There are very good introductory books on time series analysis. A very pedagogical introduction is the book by J. Brockwell and R. Davis, *Time Series: Theory and Methods* [74]. The book reviews spectral methods, autoregressive and moving average processes, state-space models and forecasting.
- A review on time series analysis applications to transcriptomic and epigenetic data is given by Z. Bar-Joseph et al. [35], with a summary of recent interesting problems and standard bioinformatic tools for analysis.
- Regarding the TDA study of the sliding window approach to series data, we recommend reading the work of J. Perea and J. Harer [404].
- For recent applications of TDA techniques to neuroscience, we recommend the review by C. Curto [130].