Simplicial Models and Topological Inference in Biological Systems

Vidit Nanda and Radmila Sazdanović

Abstract This article is a user's guide to algebraic topological methods for data analysis with a particular focus on applications to datasets arising in experimental biology. We begin with the combinatorics and geometry of simplicial complexes and outline the standard techniques for imposing filtered simplicial structures on a general class of datasets. From these structures, one computes topological statistics of the original data via the algebraic theory of (persistent) homology. These statistics are shown to be computable and robust measures of the shape underlying a dataset. Finally, we showcase some appealing instances of topology-driven inference in biological settings, from the detection of a new type of breast cancer to the analysis of various neural structures.

1 Introduction

Recent advances in genomics [40] have made it possible to sequence the entire DNA of an individual from a very small amount of that person's genetic material, say in the form of a saliva sample or a hair follicle. For each individual, one obtains as the raw output of this full sequencing process an ordered list of roughly 15 billion letters, representing the base pairs which comprise that person's DNA. This technological achievement is absolutely amazing in itself, but in all probability the bulk of its benefits will materialize over time as scientists analyze the structure of such sequences in detail. Now consider another marvel of modern engineering: the

V. Nanda
The University of Pennsylvania, Philadelphia, PA 19104, USA
e-mail: vnanda@sas.upenn.edu

R. Sazdanović (✉)
North Carolina State University, Raleigh, NC 27695, USA
e-mail: rsazdanovic@math.ncsu.edu

N. Jonoska and M. Saito (eds.), *Discrete and Topological Models in Molecular Biology*, Natural Computing Series, DOI 10.1007/978-3-642-40193-0_6,

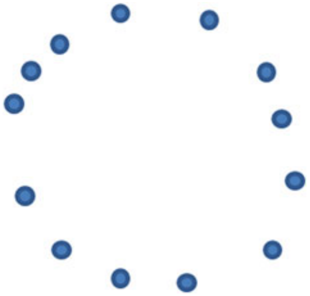

Fig. 1 A dataset consisting of points sampled from a *circle*. Although traditional line-fitting methods are likely to be uninsightful for such datasets, the methods of persistent homology can extract knowledge about the underlying shape from the point samples alone

Protein Data Bank [39] contains a wealth of structural information about protein molecules, down to the location of individual atom centers. Again, the fact that such data can now be effectively measured and collected is fascinating, but ideally one desires the ability to understand how the physical structure of a protein relates to its role in the body.

In both cases, one is confronted with enormous quantities of high-dimensional *data* prone to the usual amounts of noise or errors. From such data, one would like to extract *robust, qualitative information* and gain insight into the processes which generated the data in the first place. The standard toolkit for such inference is *statistical* at its core, and it provides computable, noise-tolerant answers to questions such as "what does the average data point look like?" or "what is the line or plane of best fit through the data?" These statistical tools are well understood, accessible to the experimentalist with a rudimentary mathematical background, and efficiently implemented in various standard software packages.

However, when the experimental data in question is produced by an essentially *nonlinear* process, the utility of our ordinary statistical tools is somewhat diminished even in the simplest of cases. Consider the dataset of Fig. 1, consisting of points sampled uniformly from a large circular figure sitting in the plane. With high probability, the average point lies near the center (but far away from the actual circle), and there is no reasonable line of best fit. It is not clear how to recover knowledge about the circle from statistics alone. Perhaps one might get lucky by noticing that the mean is roughly equidistant from all the data points, but it is easy to create slightly more complicated examples where recovering the underlying objects with any reasonable degree of accuracy from the standard statistical tools becomes hopeless. Thus, one might ask, *is there a complementary set of tools which detects the shape of the object underlying a dataset?*

A partial answer to this question comes from a previously esoteric branch of mathematics called *algebraic topology*. In particular, the theory of *persistent homology* has witnessed some success in the context of analyzing large-scale nonlinear data [16]. The basic idea behind this theory is to build an increasing family of *simplicial complexes* (indexed by a scale parameter) around the data points while carefully keeping track of the appearance and disappearance of topological features – connected components, tunnels, cavities, and their higher-dimensional cousins – as the scale parameter is increased. Numerous applications of persistent homology to various problems in the experimental sciences have been thoroughly documented elsewhere [4, 13, 18].

Although one can easily find efficient software [24, 29] for computing the persistent homology of filtered simplicial complexes, two key obstacles undermine the effective use of persistent homology to analyze experimental data. The first obstacle is an issue of *input*: how should one build a simplicial complex that captures the interesting aspects of one's data? The second issue involves the *output*: how does one make inferences about the data from the persistent homology of the input complex? With these issues in mind, the purpose of our work is threefold.

1. We provide a gentle and example-filled introduction to the mathematical *theory* which underlies (filtered) simplicial complexes. Starting with elementary combinatorial properties, we describe the connection between simplicial complexes and piecewise-linear geometry. We also discuss those algebraic objects which generate persistent homology, how they relate to simplicial geometry, and how one computes them in practice.
2. We highlight the standard *methods* of constructing filtered simplicial complexes around point cloud data via the Vietoris–Rips and Čech filtrations. We mention the relative advantages and disadvantages of these filtrations.
3. We showcase some *examples* of persistent homology in action on biological data. Recent applications have involved detection of a certain subtype of breast cancer [31] and yielded insight into the nature of neural activity – of crickets [3], monkeys, and rats [35]!

The outline of this chapter is as follows. The fundamentals of simplicial complexes and their filtrations are described in Sect. 2. Section 3 contains the core ideas needed for establishing connections between experimental data and filtrations. Section 4 describes the linear algebra of (persistent) homology and formally defines the topological features which can be detected by the theory. Finally, in Sect. 5, we survey several biological applications of persistent homology and closely related topological methods.

2 The Yoga of Simplicial Complexes

Our main goal throughout this Sect. 2 is to understand *simplicial complexes* and various related constructions. These combinatorial objects serve as a bridge between the discrete, computable world of data on one side and the continuous realm of geometric or topological spaces on the other. Our presentation here is far from complete, so we invite the interested reader to consult the wonderful texts of Munkres [28, Chaps. 1 and 2] and Spanier [36, Chap. 3] for the many gory details which we have omitted.

2.1 Simplicial Complexes

We start with a finite set V, whose elements we call *vertices*. A simplicial complex with vertex set V is a collection K of subsets of V which is closed under inclusion. More precisely, we require that the following two conditions hold:

- For each vertex v in V, the one-element set $\{v\}$ lies in K, and
- If τ is in K and $\sigma \subset \tau$ is a subset, then σ is also in K.

Each element τ of K is called a *simplex*, and its *dimension* (written $\dim \tau$) is defined to be $\#(\tau) - 1$, where $\#$ denotes the cardinality (i.e., it counts the number of vertices of τ). Any subset σ of τ is called a *face* of τ, and this relationship is denoted by $\sigma \preceq \tau$. We write K_d to indicate the collection of d-dimensional simplices in K for each $d \geq 0$. It is clear from the first property of simplicial complexes that the elements of V correspond in a one-to-one manner with those of K_0, and it is therefore customary to speak of the two sets interchangeably. Consequently, one often encounters phrases resembling “let K be a simplicial complex” with no explicit mention of the underlying vertex set. Before proceeding any further, we will examine a small simplicial complex in some detail.

Example 1. Given a vertex set $V = \{a, b, \ldots, f, g\}$, we may construct a simplicial complex K in layers, one dimension at a time. We denote subsets of V by their elements in alphabetical order, so that $\{a, b, c\}$ is simply written *abc*. As we have already seen, K_0 is completely determined by V. Next, K_1 can contain any pair of distinct vertices in V and there is some freedom to choose such pairs. For instance, we can select

$$K_1 = \{ab, ac, ae, bc, bd, be, bg, cd, cg, dg, ef\}.$$

Fixing K_1 immediately constrains which simplices can lie in K_2. For instance, *abe* is allowed in K_2, since all of its one-dimensional faces *ab*, *ae*, *be* are in K_1. However, *acd* is banned because $ad \prec acd$ but *ad* is not present in K_1. We add the following (legal!) two-dimensional simplices to K:

$$K_2 = \{abe, bcg, bcd, bdg, cdg\},$$

and note that the only three-dimensional simplex whose faces all exist in K_2 is *bcdg*. Let us include that simplex as well, and we obtain

$$K_3 = \{bcdg\}.$$

No four-dimensional simplices are allowed, since the presence of a single such simplex would require K_3 to have at least four elements, so our K is just the union of K_d for d in $\{0, 1, 2, 3\}$. This complex K is reasonably small and low-dimensional; it is often useful to visualize such simplicial complexes as embedded in Euclidean space (see Fig. 2).

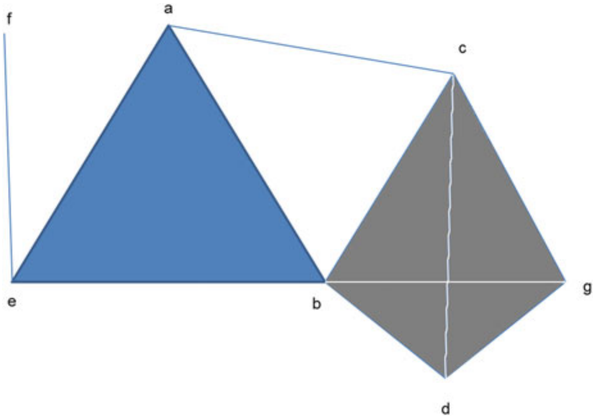

Fig. 2 A pictorial representation of the simplicial complex K of Example 1 with points representing vertices. The *lines*, *triangles*, and tetrahedra stand in for one, two, and three-dimensional simplices, respectively. We note that K consists of a single connected component and that the 1-simplices ab, ac, bc form a loop

2.2 Subcomplexes, Filtrations, and Sublevelsets

Let K be any simplicial complex. A subcollection L of simplices from K which forms a simplicial complex in its own right is called a *subcomplex* of L, written $L \hookrightarrow K$. In other words, *if a simplex τ lies in L, then all of its faces in K are also present in L*. In general, the vertex set of L may be strictly smaller than that of K, with equality only occurring when $L_0 = K_0$. The reader may enjoy proving the following result, but we have our doubts.

Proposition 1. *If simplicial complexes K, L, and M satisfy $L \hookrightarrow K$ and $K \hookrightarrow M$, then we also have $L \hookrightarrow M$.*

Let $N \geq 1$ be a natural number and K a simplicial complex. A *filtration* $\mathscr{F}$ of the simplicial complex K is a nested collection of subcomplexes $\mathscr{F}_n K \hookrightarrow K$ for n in $\{0, \ldots, N\}$ which ascends from the empty set $\emptyset$ all the way up to K like this:

$$\emptyset = \mathscr{F}_0 K \hookrightarrow \mathscr{F}_1 K \hookrightarrow \mathscr{F}_2 K \hookrightarrow \cdots \hookrightarrow \mathscr{F}_{N-1} K \hookrightarrow \mathscr{F}_N K = K.$$

Here, N is called the *length* of $\mathscr{F}$. The simplicial complex K trivially forms a length-1 filtration, since we have $\emptyset \subset K$. A slightly less obvious filtration could be constructed by dimension: let $\mathscr{F}_n K$ be the collection of all simplices of dimension at most n. But we will consider a more interesting example. In particular, we would like to illustrate the fact that the process of building K from subcomplexes along $\mathscr{F}$ causes various interesting intermediate features to appear and disappear.

Example 2. Let K be the simplicial complex of Example 1. We will define a filtration $\mathscr{F}$ of K which has length 4 by describing each subcomplex $\mathscr{F}_n K$

individually. Since $\mathscr{F}_0 K$ is empty, we ignore it and move on to the first subcomplex,

$$\mathscr{F}_1 K = \{a, b, c, d, f, ac, cd, eb\}.$$

There are three pieces in this subcomplex, as the first quarter of Fig. 3 reveals. Next, we consider

$$\mathscr{F}_2 K = \mathscr{F}_1 K \cup \{g, ab, ae, bc, bd, ef, bcd\},$$

where $\cup$ indicates a union of sets. The addition of the vertex g adds yet another piece to the three already present in $\mathscr{F}_1 K$, but ab and ef join three of those pieces into a single large component. The sequences (ab, ae, be) and (ab, ac, bc) of one-dimensional simplices form two *loops*. A similar loop formed by (bc, bd, cd) is immediately filled by the two-dimensional simplex bcd. Moving on, we define

$$\mathscr{F}_3 K = \mathscr{F}_2 K \cup \{abe, bcg, bdg, cdg\}.$$

The simplex abe fills up the loop (ab, ae, be) consisting of its faces. The addition of the other simplices reveals a new feature: a void, or *cavity*, formed by (bcd, bcg, bdg, cdg). This cavity is very different from the loops that we have encountered before, in the sense that – at least as pictured in Fig. 3 – it encloses a three-dimensional region rather than a planar one. Finally, we add

$$\mathscr{F}_4 K = \mathscr{F}_3 K \cup \{bcdg\},$$

and this last simplex fills the cavity obtained from $\mathscr{F}_3 K$.

Let K be any simplicial complex, and let $\mathbf{N}$ denote the natural numbers. Consider a function $g : K \to \mathbf{N}$ which assigns to each simplex σ a natural number $g(\sigma)$. Then, the *sublevelset* of g at the natural number n is defined by $S_n(g) = \{\sigma \in K \mid g(\sigma) \leq n\}$. Clearly, we have $S_n(g) \subset S_{n+1}(g)$ as sets. Unfortunately, $S_n(g)$ is not always a subcomplex of K for arbitrary functions g: if $g(\sigma) > n \geq g(\tau)$ with $\sigma \prec \tau$, then $S_n(g)$ contains τ but not its face σ. It turns out that this is the only obstruction to having a filtration by sublevelsets, so we will restrict our choice of g to functions which avoid this behavior.

We call $g : K \to \mathbf{N}$ *monotone* if it increases with dimension along faces. Thus, g is monotone (or *order-preserving*) if $g(\sigma) \leq g(\tau)$ whenever $\sigma \preceq \tau$. In this case, it is easy enough to check that setting

$$\mathscr{F}_n K = S_n(g) = \{\tau \in K \mid g(\tau) \leq n\}$$

yields a filtration of K whose length equals the maximum number of distinct values attained by g on K. One could also consider monotone maps $g : K \to \mathbf{R}$ to the real numbers, but this would be largely for convenience. Since there are only finitely many simplices in K, the image of g may assume only finitely many distinct real

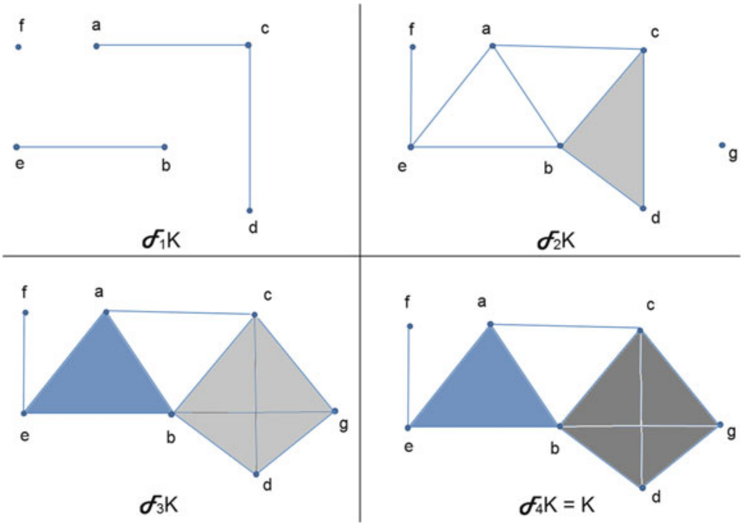

Fig. 3 An illustrated view of the filtration $\mathscr{F}$ defined in Example 2. Note that the intermediate stages of the filtration look very different from K! A systematic study of the appearance and disappearance of features such as connected components, loops, and cavities provides a coarse-grained view of how K is incrementally built along its subcomplexes in $\mathscr{F}$. See Example 2 for details

values. Indexing these values $\{c_1, \ldots, c_N\}$ in ascending order yields a one-to-one monotone correspondence with a subset of $\mathbf{N}$: just send c_n to n. Thus, an $\mathbf{R}$-valued g can easily be replaced by an $\mathbf{N}$-valued cousin with no essential change in the structure of the sublevelset filtration. Sublevelset filtrations are ubiquitous for the following simple reason.

Proposition 2. *For any filtration $\mathscr{F}$ of a simplicial complex K, there is a unique monotone function $g : K \to \mathbf{N}$ such that $\mathscr{F}$ is the sublevelset filtration of g.*

The proof is easy: let g be the function that sends each simplex σ in K to the smallest n such that σ is a simplex in $\mathscr{F}_n K$. The reader may wish to check, for example, that setting $g(\sigma) = \dim(\sigma)$ retrieves the filtration by dimensions mentioned before Example 2.

2.3 The Geometry of Simplices: Realizations and Simplicial Maps

As we have remarked before, a primary advantage of simplicial complexes is their ability to interface between discrete and continuous spaces. When we visualize simplicial complexes (see Fig. 2, for example), we use nondiscrete geometric objects such as lines, triangles, and tetrahedra. The reader may have noticed that much of the terminology for simplicial complexes (for instance "dimension" and "vertex") appears to have been borrowed from corresponding notions for these familiar and concrete geometric objects. There is a standard protocol underlying this dictionary between simplices and these objects, which we will now describe. All that is assumed of the reader is a basic understanding of d-dimensional Euclidean real space $\mathbf{R}^d$, each point of which consists of an ordered sequence of d real numbers. For each j between 1 and d, the j-th *basis vector* $\mathbf{e}_j$ of $\mathbf{R}^d$ is identified with the point which contains a 1 in the j-th component and 0's everywhere else.

We fix a dimension d, and let $\mathbf{u} = \{\mathbf{u}_1, \ldots, \mathbf{u}_M\}$ be a collection of $M \geq 1$ points in $\mathbf{R}^d$. A *convex combination* of these points is any point in $\mathbf{R}^d$ which can be expressed as an $\mathbf{R}$-linear combination

$$x = p_1\mathbf{u}_1 + \cdots + p_M\mathbf{u}_M,$$

where each coefficient p_m is nonnegative and the sum $p_1 + \cdots + p_M$ of all these coefficients equals 1. The *convex hull* of this collection $\mathbf{u}$ is the set of all such convex combinations,[1] and we denote this subset of $\mathbf{R}^d$ by $\mathrm{Conv}(\mathbf{u})$. Now let $\mathbf{v} = \{\mathbf{v}_1, \ldots, \mathbf{v}_N\}$ be another collection of points in $\mathbf{R}^d$, and assume that we are given a map $\eta : \mathbf{u} \to \mathbf{v}$. Then, η provides a standard and fairly obvious recipe for concocting a map $\bar{\eta} : \mathrm{Conv}(\mathbf{u}) \to \mathrm{Conv}(\mathbf{v})$ as follows:

$$\bar{\eta}(p_1\mathbf{u}_1 + \cdots + p_M\mathbf{u}_M) = p_1\eta(\mathbf{u}_1) + \cdots + p_M\eta(\mathbf{u}_M).$$

If we restrict our attention to the subcollection $\mathbf{u}'$ of $\mathbf{u}$ and let η be the inclusion map $\mathbf{u}' \to \mathbf{u}$, we immediately see that $\mathrm{Conv}(\mathbf{u}') \subset \mathrm{Conv}(\mathbf{u})$. The d-dimensional *standard simplex* $\Delta^d \subset \mathbf{R}^{d+1}$ is defined to be $\mathrm{Conv}(\mathbf{e}_1, \ldots, \mathbf{e}_{d+1})$, the convex hull of the basis elements of $\mathbf{R}^{d+1}$. Equivalently,

$$\Delta^d = \left\{(x_1, \ldots, x_{d+1}) \mid \text{each } x_j \geq 0 \text{ and } x_0 + \cdots + x_{d+1} = 1\right\}.$$

For example, Δ^2 is the two-dimensional triangle determined by the standard basis vectors $\mathbf{e}_1 = (1, 0, 0)$, $\mathbf{e}_2 = (0, 1, 0)$ and $\mathbf{e}_3 = (0, 0, 1)$ in three-dimensional Euclidean space.

Let K be a simplicial complex with $d + 1$ vertices, which we order as $\{v_1, \ldots, v_{d+1}\}$. We will construct a concrete subset $|K|$ of the standard simplex Δ^d, which is the canonical geometric space associated to K. For each simplex

[1] It is easy to work out that the convex hull of two points is the line segment connecting them, and that the convex hull of three points (which do not all lie on the same line) is the triangle containing those three points as vertices. In higher dimensions and with many more points, things become less obvious. Determining convex hulls is a fundamental problem in computational geometry.

σ in K consisting of vertices $\{v_{i_1}, \dots, v_{i_m}\}$, we first define $|\sigma| \subset \Delta^d$ by $|\sigma| = \text{Conv}(\mathbf{e}_{i_1}, \dots, \mathbf{e}_{i_m})$.

Definition 1. The *geometric realization* $|K| \subset \Delta^d$ of K is the union of all $|\sigma|$ as σ ranges over simplices in K.

Thus, a counterpart to K has been delineated within Δ^d as a concrete geometric object. Before being completely satisfied with this definition, however, one might wonder: *what happens if we order the vertices* $\{v_1, \dots, v_{d+1}\}$ *differently?* In order to arrive at a satisfactory answer, we must understand when two simplicial complexes are considered equivalent; for this purpose, we turn our attention to *simplicial maps*. These maps will require a domain and a range, so let K and L be simplicial complexes with vertex sets U and V, respectively.

Definition 2. A *simplicial map* $\phi : K \to L$ assigns to each vertex u in U a vertex $\phi(u)$ in V so that the image of each simplex $\sigma \in K$ constitutes a simplex $\phi(\sigma) \in L$.

Here, by $\phi(\sigma)$ we mean the set of vertices in V obtained by mapping each vertex of σ by ϕ into V. We have already seen examples of simplicial maps: if $K \hookrightarrow L$, then the map sending each vertex of K to itself as a vertex of L is simplicial. It turns out that any simplicial map $\phi : K \to L$ induces a continuous function of geometric realizations, which we denote by $|\phi| : |K| \to |L|$. This map acts exactly as one would expect. Namely, we note first that $|K|$ is a union of realizations of simplices $|\sigma|$ where $\sigma \in K$, so it suffices to understand how each individual $|\sigma|$ is mapped by $|\phi|$. Since ϕ maps the vertices of σ into the vertices of its image $\phi(\sigma)$, the map $\bar{\phi}$ linearly maps the convex hull $|\sigma|$ into the convex hull $|\phi(\sigma)|$. We define the function $|\phi|$ to be that transformation from $|K|$ to $|L|$ which acts on each $|\sigma| \subset |K|$ as the linear map $\bar{\phi}$. Thus, although $|\phi| : |K| \to |L|$ itself may not be a linear map, its action on each convex piece $|\sigma|$ of $|K|$ is linear. For this reason, the continuous maps between realizations induced by simplicial maps are often called *piecewise-linear* maps.

The reader is warned that arbitrary simplicial maps do not preserve dimension: one might have $\dim \phi(\sigma) < \dim \sigma$ if ϕ is not one-to-one on the vertices of σ. On the other hand, if ϕ is a *bijection* – a map that associates each vertex of U to a single vertex of V and vice versa – then not only are dimensions preserved, but also the net effect of mapping K into L via ϕ is essentially that of relabeling the vertices. In such a case, K and L are called *isomorphic* and we write $K \simeq L$. Although the geometric realizations $|K|$ and $|L|$ might disagree in terms of exactly how they sit in $\mathbf{R}^d$, they are topologically (and indeed, geometrically) equivalent because an invertible linear transformation of $\mathbf{R}^d$ (i.e., an invertible matrix) maps $|K|$ to $|L|$, with its inverse taking $|L|$ back into $|K|$. More precisely, any simplicial map $\phi : K \to L$ induces a map from the basis of $\mathbf{R}^{\#U}$ to that of $\mathbf{R}^{\#V}$ as follows:

$$\text{basis element} \overset{\simeq}{\longleftrightarrow} \text{vertex of } K \xrightarrow{\phi} \text{vertex of } L \overset{\simeq}{\longleftrightarrow} \text{basis element}.$$

Following this diagram from left to right produces a matrix which maps $\Delta^{\#U-1}$ to $\Delta^{\#V-1}$ so that the image of $|K|$ is contained inside $|L|$. In the special case where ϕ is a bijection of vertices, this matrix is invertible. It is in this sense that simplicial complexes (up to equivalence by isomorphism) are uniquely associated with their geometric realizations (up to equivalence by invertible linear transformations).

3 Constructing Filtrations Around Points

The process of conducting experiments and collecting data is, by its very nature, the crux of all experimental science. Experimental data can take many forms, including text, images, and even video. For our purposes, we will restrict our attention to a very specific form of data: a *point cloud*. By a point cloud, we simply mean a finite collection P of points in $\mathbf{R}^d$ for some suitable dimension d, and make no further assumptions regarding the nature of P. We would like to remark here that it is possible to construct faithful point cloud representations of just about any type of data, although it may not be advantageous to do so because the dimension d might become enormous.

A first step towards applying topological machinery to a point cloud is to construct a filtration of a simplicial complex whose vertex set can in some way be identified with P. We will discuss two standard filtrations that may be constructed around point clouds. Along the way, we will try to highlight their relative advantages and disadvantages.

The largest possible simplicial complex with vertex set P is, of course, the *complete* simplicial complex, where every possible subset of P constitutes a simplex. We will denote this complex by K_P throughout this section.[2] All the filtrations that we encounter here will be – either implicitly or explicitly – filtrations of K_P.

There are many notions of *distance* that one can reasonably impose on $\mathbf{R}^d$. For any $p \geq 1$, we can consider the p-distance

$$\mathbf{d}_p(x, y) = \sqrt[p]{\sum_{m=1}^{d} |x_m - y_m|^p},$$

so that the familiar Euclidean distance is recovered when one sets $p = 2$. Another option is the max-distance,

$$\mathbf{d}_\infty(x, y) = \max_{1 \leq m \leq d} \{|x_m - y_m|\}.$$

[2] In fact, K_P consists of a single $(\#P - 1)$-dimensional simplex along with all its faces!

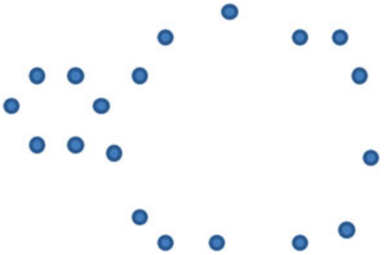

Fig. 4 An example of a small point cloud sitting in two-dimensional Euclidean space. Note that the points appear to have the shape of two *circles*, one larger than the other

The filtrations that one constructs around $P \subset \mathbf{R}^d$ depend on which notion of distance is chosen. In order to provide the most flexibility, we will simply denote the distance we use by $\mathbf{d}$ and leave the explicit choice to the reader.

For any positive real number $r \geq 0$ and a point $x \in \mathbf{R}^d$, we define the *ball of radius r around x* as

$$B_r(x) = \left\{ y \in \mathbf{R}^d \mid \mathbf{d}(x, y) < r \right\}.$$

The shape of this ball depends on the distance $\mathbf{d}$. The reason for calling this type of set a “ball” becomes clear when one uses the standard distance $\mathbf{d}_2$.

As a running example, we will consider a toy example of a point cloud in $\mathbf{R}^2$ as shown in Fig. 4. The exact coordinates of each point are relatively unimportant; we are only seeking *qualitative* information. Thus, what we will focus on here is the fact that the point cloud appears to contain two distinct loops, with the one on the right-hand side having a larger diameter than the other. In our running example, we will use the distance $\mathbf{d}_2$.

3.1 The Vietoris–Rips Filtration

Let $P \subset \mathbf{R}^d$ be our point cloud. One can compute all the pairwise distances $\mathbf{d}(p, p')$ between pairs of points p and p' in P. This data structure – consisting of P along with the pairwise distances – suffices to construct the Vietoris–Rips filtration (Fig. 5). At any given *scale* $\epsilon \geq 0$, we define the simplicial subcomplex $\mathcal{V}_\epsilon K_P$ of the complete complex K_P as follows. The vertex set is P, and each simplex σ in $\mathcal{V}_\epsilon K_P$ consists of a subcollection of vertices so that the pairwise distance between any two is less than ϵ. Let $\sigma \subset P$ be a subcollection of points $(p_1, \ldots, p_m)$. Restricting the indices i and j to $\{1, \ldots, m\}$, we have

$$\sigma \text{ is a simplex in } \mathcal{V}_\epsilon K_P \text{ if } \mathbf{d}(p_i, p_j) < \epsilon \text{ for all } i, j,$$

or equivalently,

$$\sigma \text{ is a simplex in } \mathcal{V}_\epsilon K_P \text{ if } B_{\epsilon/2}(p_i) \cap B_{\epsilon/2}(p_j) \neq \emptyset \text{ for all } i, j.$$

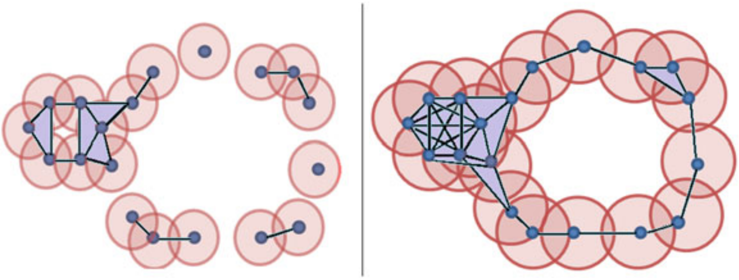

Fig. 5 Two stages of the Vietoris–Rips filtration around the point cloud from Fig. 4. The scale ϵ increases from *left* to *right*, and the balls of radius ϵ have been shown underlying the simplices. The smaller loop is captured faithfully at the smaller ϵ value, and the larger loop is captured at the larger ϵ value. But *no single scale captures both*!

Here $\cap$ stands for the intersection of sets.

It is easy to see that for any value of ϵ, our definition of $\mathscr{V}_\epsilon$ yields a genuine simplicial complex. After all, if σ is a simplex and τ is a face of σ, then the set of all pairwise distances between vertices of τ is contained in the set of the corresponding pairwise distances of σ's vertices. On the other hand, we can also immediately check that for $\delta > \epsilon$, we have $\mathscr{V}_\epsilon K_P \hookrightarrow \mathscr{V}_\delta K_P$ because if all pairwise distances are less than ϵ, they are also less than δ.

We define the function $g_{\mathscr{V}} : K_P \to \mathbf{R}$ as follows. For any simplex σ in K_P,

$$g_{\mathscr{V}}(\sigma) = \max_{p,q \text{ in } \sigma} \{\mathbf{d}(p,q)\}.$$

Whenever $\sigma \prec \tau$, we obtain $g_{\mathscr{V}}(\sigma) \leq g_{\mathscr{V}}(\tau)$ because we are taking the maximum over a larger set. Thus, g is monotone, and the following definition makes sense by Proposition 2.

Definition 3. The *Vietoris–Rips filtration* around $P \subset \mathbf{R}^d$ is the sublevelset filtration of $g_{\mathscr{V}}$.

We place the pairwise distances between points in P in ascending order, $0 \leq \epsilon_1 \leq \cdots \leq \epsilon_N$, and note that we have

$$\mathscr{V}_{\epsilon_1} K_P \hookrightarrow \mathscr{V}_{\epsilon_2} K_P \hookrightarrow \cdots \hookrightarrow \mathscr{V}_{\epsilon_N} K_P = K_P.$$

In practice, one stops well short of constructing the Vietoris–Rips filtration all the way up to ϵ_N, unless the number of points in P is very small. The reason for this is simple: the complete complex K_P contains as many simplices as there are nonempty subsets of P, so its cardinality is $2^{\#P} - 1$. Even for a tiny cloud containing only 40 points, building the full Vietoris–Rips filtration requires storing well over a *trillion* simplices in system memory!

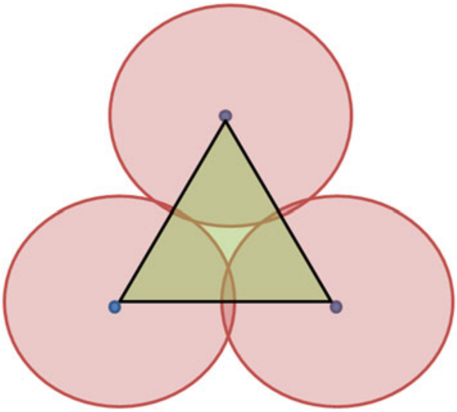

Fig. 6 A collection of balls in the plane with nonempty pairwise intersection but no triple intersection. The union of these balls clearly encloses a hole, which the overlaid Vietoris–Rips filtration fails to capture at the current radius

Advantages. Pairwise distances are easily *computable* in most settings, so, at least in principle, it is very easy to determine the scale at which a given simplex joins the Vietoris–Rips filtration. Since one only requires knowledge of pairwise distances, this filtration is extremely *flexible* in the sense that one can construct it around extremely general data types. For instance, consider a situation where the data arises from measuring correlations between various states of a complex system. In this case, it may not be natural to try to embed these states as points in some $\mathbf{R}^d$. However, a knowledge of the pairwise correlations alone is enough to construct the Vietoris–Rips complex!

Disadvantages. As we have already discussed, the Vietoris–Rips filtration is liable to become *gigantic* in terms of the number of simplices because its size scales exponentially with the number of points. Moreover, there is no control over the *dimensions* of simplices that are built, even for small values of the scale ϵ: if there are 20 points with pairwise distances all less than ϵ, then the 19-dimensional simplex containing those points will belong to $\mathscr{V}_\epsilon K_P$ even if those points are sitting in two-dimensional space! A subtler issue with these filtrations is that they are merely *approximations* to the structure of the underlying space which do not recover its structure accurately at each scale ϵ. It is easy to construct – at least with the distance $\mathbf{d}_2$ – three balls so that any pair intersects, but there is no common point in the intersection of all three, thus forming a hole (see Fig. 6). However, the geometric realization of the resulting Vietoris–Rips filtration at the given scale fails to capture that hole, because it contains the two-dimensional simplex spanning the ball centers.

A description of efficient algorithms for constructing Vietoris–Rips filtrations may be found in [41]. Most persistent-homology software packages (e.g., [29]) contain implementations of these algorithms.

3.2 The Čech Filtration

Letting $P \subset \mathbf{R}^d$ be our point cloud and K_P the complete simplicial complex with vertex set P, we define a simplicial subcomplex $\mathscr{C}_\epsilon K_P$ of K_P at each scale $\epsilon > 0$ in the following way. A subcollection $\sigma \subset P$ of points forms a simplex of $\mathscr{C}_\epsilon K_P$ if there exists some point x in $\mathbf{R}^d$ whose distance[3] from each vertex of σ is at most ϵ. More precisely, let $\sigma = (p_1, \ldots, p_m)$. Then,

$$\sigma \text{ is a simplex in } \mathscr{C}_\epsilon K_P \text{ if } \mathbf{d}(x, p_i) < \epsilon \text{ for all } i \text{ and some fixed } x,$$

or, equivalently,

$$\sigma \text{ is a simplex in } \mathscr{C}_\epsilon K_P \text{ if the intersection } \bigcap_{j=1}^{m} B_\epsilon(p_i) \neq \emptyset.$$

It is apparent that the construction of Čech filtrations depends crucially on the following computation: *given a collection σ of points in $\mathbf{R}^d$, what is the smallest radius r so that there exists some point x in $\mathbf{R}^d$ whose distance from each point in σ is less than r?* Discrete and computational geometers often refer to this as the *smallest enclosing ball* problem: after all, the ball of radius r around x encloses all the points in σ and, by definition, it must be the smallest ball to do so. Although there are various algorithms available to compute this minimal ball (some sacrifice exactness for speed), in general (for large point sets sitting in high dimensions) this is a complicated, nontrivial problem. Certainly, one requires a lot more computational muscle than the simple pairwise distance calculations that must be performed for constructing a Vietoris–Rips filtration.

Consider the function $g_\mathscr{C} : K_P \to \mathbf{R}$ defined on the simplex $\sigma = (p_1, \ldots, p_m)$ by

$$g_\mathscr{C}(\sigma) = \min\left\{r \geq 0 \mid \text{ there is an } x \text{ in } \mathbf{R}^d \text{ with } \mathbf{d}(x, p_j) < r \text{ for } 1 \leq j \leq m\right\}.$$

It is clear that $g_\mathscr{C}$ is monotone; if $\tau \prec \sigma$ and some ball $B_r(x)$ contains all the vertices of σ, then it also contains the subset of vertices which belong to τ, and hence the smallest enclosing ball for τ can have radius no larger than r.

Definition 4. The *Čech filtration* around $P \subset \mathbf{R}^d$ is the sublevelset filtration of $g_\mathscr{C}$.

Since there are only finitely many simplices in K_P, this function $g_\mathscr{C}$ assumes only finitely many values. Listing them in increasing order as $0 = \epsilon_0 \leq \epsilon_1 \leq \cdots \leq \epsilon_N$, one obtains

[3] This x is not necessarily a point in the cloud P, so typically the Čech filtration cannot be built from knowledge of pairwise distances alone!

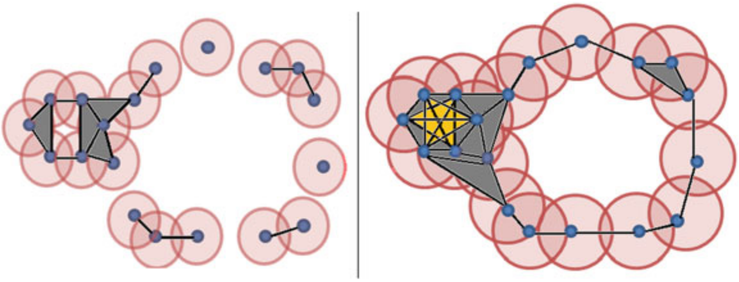

Fig. 7 Two stages of the Čech filtration of the point cloud of Fig. 4. Note that there are fewer simplices of dimension 2 and above when compared with Fig. 5 at the larger scale. For instance, the two differently colored simplices are *not* present in the Čech filtration, although they are present in the Vietoris–Rips filtration at the same scale

$$\mathscr{C}_{\epsilon_1} K_P \hookrightarrow \mathscr{C}_{\epsilon_2} K_P \hookrightarrow \cdots \hookrightarrow \mathscr{C}_{\epsilon_N} K_P = K_P.$$

The fact that higher-order intersections are taken into account when one is building the Čech filtration incurs a computational burden, but there is a substantial payoff. In particular, it follows from a result known as the *nerve theorem* that the geometric realization $|\mathscr{C}_\epsilon K_P|$ is topologically equivalent[4] to the union of all balls $B_\epsilon(p)$, where p ranges over the points in P.

Advantages. At each scale ϵ, the Čech filtration is *faithful* to the topology of the union of balls. In particular, keeping track of higher-order intersections allows one to bypass the issue highlighted in Fig. 6: the two-dimensional simplex concerned will not enter the Čech filtration until the scale where all three balls intersect, at which point there is no loop. For the same reason, the Čech filtration at a given scale is typically much *smaller*, in the sense that it contains fewer simplices than the Vietoris–Rips filtration at the same scale: see Fig. 7.

Disadvantages. As we have already noted, the *complexity* of the enclosing ball problem makes it difficult to construct Čech filtrations around point clouds in dimensions exceeding 3. As with the Vietoris–Rips filtration, a cluster of nearby points produces a simplex of high *dimension* regardless of the ambient dimension d.

Algorithms to construct the Čech filtration are described in [11], and an implementation is available as part of [24].

[4] This equivalence is up to a fundamental topological invariant known as *homotopy*.

3.3 Other Filtrations

While the two filtrations mentioned above are the most common ones encountered in practice, there are several other approaches to imposing simplicial structures on point clouds. In particular, the *witness complex* filtration [12] attempts to reduce the number of simplices by preprocessing the point cloud P itself as follows. We fix an acceptable "fuzz" parameter $\delta > 0$, and restrict our attention to a subset $P' \subset P$ of *landmark* points so that no two are within δ of each other. This preprocessing allows us to reduce the dimension (and hence the number) of simplices which appear at each scale $\epsilon > \delta$ in either the Vietoris–Rips or the Čech filtration. This computational advantage is not without a price, however: to the best of our knowledge, there are no explicit results about how faithfully a witness complex represents the topology of the underlying union of balls.

A drastically different approach, which is especially useful in low dimensions, involves the use of filtered *alpha complexes* [15]. Although these complexes require even more computational-geometry muscle to construct than the Čech filtration, the benefits are immense. The nerve theorem applies in the context of alpha complexes, so they are also topologically faithful to the underlying union of balls, like Čech filtrations. At the same time, the dimension of the simplices encountered in an alpha complex never exceeds d, the ambient Euclidean dimension!

4 Homology and Its Computation

Throughout the preceding sections, we have discussed various topological *features* – such as loops and cavities – which appear in geometric realizations of simplicial complexes or in the context of point clouds thickened into balls by some scale ϵ. In order to precisely understand the objects which encode and catalog such features, we must turn to algebra. Any reader who experiences moral qualms about our descent from the Olympus of geometric shapes to the Hades of algebraic formalism stands in distinguished company:

> Algebra is the offer made by the devil to the mathematician. The devil says: "I will give you this powerful machine, it will answer any question you like. All you need to do is give me your soul: give up geometry and you will have this marvellous machine."
>
> Sir Michael Atiyah

4.1 The Linear Algebra of Holes

The "marvellous machine" called *homology* detects "holes" of all dimensions by using linear algebra. It associates to each simplicial complex K a collection of algebraic objects $\mathsf{H}_d(K)$ called *homology groups*, where d ranges over the dimensions of the simplices encountered in K. Given a simplicial map $\phi : K \to L$, homology produces *group homomorphisms* $\phi_d^* : \mathsf{H}_d(K) \to \mathsf{H}_d(L)$. The type of groups and homomorphisms that one obtains depends on the choice of some underlying *coefficient system*. Here, we will use the real numbers **R**. In this setting, each homology group is just some Euclidean space and each homomorphism a matrix with entries in **R**.

Let K be a simplicial complex with ordered vertices. What this means for our purposes is that the vertices of any simplex σ can be uniquely written in some ascending order $(v_0, \dots, v_d)$. The d-dimensional *chain group* $\mathsf{C}_d(K)$ of K consists of **R**-linear combinations of d-dimensional simplices. Thus, a typical element of $\mathsf{C}_d(K)$ – called a d-dimensional *chain* – is $a_1\sigma_1 + \cdots + a_m\sigma_m$, where the a's are real numbers and the σ's are d-dimensional simplices. Clearly, this chain group is equivalent to $\#K_d$-dimensional Euclidean space: just use the d-dimensional simplices as a basis. Let $\sigma = (v_0, \dots, v_d)$ be such a basis element, and for each j in $\{0, \dots, d\}$ let σ_j be that $(d-1)$-dimensional proper face of σ which contains all the vertices except v_j. Now, the *boundary* of σ is a $(d-1)$-dimensional chain given by the alternating sum of these faces:

$$\partial_d(\sigma) = \sigma_0 - \sigma_1 + \cdots + (-1)^d \sigma_d.$$

Thus, ∂_d defines a linear transformation $\mathsf{C}_d(K) \to \mathsf{C}_{d-1}(K)$, and hence may be thought of as a matrix once we order the simplices into a basis. We define the d-dimensional *cycle group* $\mathsf{Z}_d(K)$ to be the subspace corresponding to the kernel of this matrix in $\mathsf{C}_d(K)$, and the $(d-1)$-dimensional *boundary group* $\mathsf{B}_{d-1}(K)$ is the image of this matrix as a subspace of $\mathsf{C}_{d-1}(K)$. The elements of $\mathsf{Z}_d(K)$ and $\mathsf{B}_d(K)$ are called the d-dimensional *cycles* and *boundaries*, respectively. It can be checked that each d-dimensional boundary is also a cycle.[5] Now, the d-dimensional *homology group* is defined as the quotient

$$\mathsf{H}_d(K) = \frac{\mathsf{Z}_d(K)}{\mathsf{B}_d(K)}.$$

Thus, we are interested in cycles, but do not distinguish between two cycles if they are related by a boundary. That is, we *partition* the cycles x from $\mathsf{Z}_d(K)$ into *homology classes* $[x]$, with $[x] = [y]$ whenever $x - y$ lies in $\mathsf{B}_d(K)$.

To see why we care about this quotient, let us go back to the complex of Example 1. Observe that the loop formed by ab, ac, bc corresponds to the algebraic cycle $x = ab + bc - ac$, whose boundary is 0. So far, so good. But, algebraically, even ab, ae, be forms a "loop": let $y = ab + be - ae$, and check that $\partial_1(y) = 0$. The difference between these cycles – transparent to the eye but opaque to the algebra at this point – is the presence of abe which fills up the latter cycle. In order to make

[5] To see why this is the case, note that the composition $\partial_d \circ \partial_{d+1}$ is the zero map from $\mathsf{C}_{d+1}(K)$ to $\mathsf{C}_{d-1}(K)$ for each dimension d.

the chain algebra recognize this fill-up, we note that $\partial_2(abc) = ab + bc - ac$. In the quotient space, this cycle y therefore ends up in the trivial homology class [0]. This is why algebraic cycles alone are not enough; we need to quotient by the boundaries of higher simplices.

4.2 Smith Normal Form and Betti Numbers

Staying with the simplicial complex K of Example 1, let us see what it takes to compute $\mathsf{H}_0(K)$. First, we order the zero- and one-dimensional simplices of K in some consistent way. For convenience, we may choose the alphabetical order, and hence obtain the following sequences of cells:

$$K_0 = (a, b, c, d, e, f) \text{ and } K_1 = (ab, ac, ae, bc, bd, be, bg, cd, cg, dg, ef).$$

Next, we express the boundary operator $\partial_1 : C_1(K) \to C_0(K)$ as a matrix M_1 in our chosen basis. For instance, in the column for ac and the row for a, one finds the dot product $\langle \partial_1(ac), a\rangle = -1$, which simply extracts the coefficient of a in the boundary of ac. Proceeding in this fashion yields the following matrix:

$$M_1 = \begin{array}{c|ccccccccccc|} & ab & ac & ae & bc & bd & be & bg & cd & cg & dg & ef \\ \hline a & -1 & -1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ b & 1 & 0 & 0 & -1 & -1 & -1 & -1 & 0 & 0 & 0 & 0 \\ c & 0 & 1 & 0 & 1 & 0 & 0 & 0 & -1 & -1 & 0 & 0 \\ d & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & -1 & 0 \\ e & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & -1 \\ f & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ g & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 \end{array}.$$

Using standard row and column operations (with coefficients in $\mathbf{R}$), we can put M_1 in *Smith normal form*, so that the off-diagonal entries are all zero, and the diagonal contains only zeros and ones. The number of zero entries in the diagonal of the Smith normal form is then equal to the *rank* of $H_0(K)$ as a vector space over the real numbers. One can repeat this process for all dimensions $d \geq 1$: order the cells, generate a matrix representation M_d of ∂_d in the chosen basis, and compute its Smith normal form. The number of zero entries in the diagonal of M_d's Smith normal form is called the $(d-1)$-th *Betti number* of K, and it equals the rank of $\mathsf{H}_{d-1}(K)$ as a Euclidean space. Keeping track of the change-of-basis matrices of the row and column operations also produces an explicit basis for $\mathsf{H}_{d-1}(K)$ in terms of the chains in $\mathsf{C}_{d-1}(K)$.

One may ask: *what does it all mean?* The answer is easy in low dimensions: the zero-, one-, and two-dimensional Betti numbers count the *connected components,* *tunnels, and cavities*, respectively, of the underlying simplicial complex.[6] In higher dimensions, the answer is subtler because we lose the ability to visualize geometry. But in any case, the Betti numbers of a simplicial complex provide computable *topological statistics* of that complex. The structure encoded by the actual groups (not just the Betti numbers) is much more intricate, but it should be clear (at least in principle) that knowledge of those groups as quotients of chains enables one to actually find components, tunnels, cavities, and their higher-dimensional analogs as linear combinations of simplices.

The situation is very similar for simplicial maps $\phi : K \to L$. Since ϕ sends simplices of K to simplices of L, for each dimension d it induces a *chain map* $\phi_d^\# : \mathsf{C}_d(K) \to \mathsf{C}_d(L)$ determined by the following action on the basis elements. Given $\sigma \in K_d$, we define

$$\phi_d^\#(\sigma) = \begin{cases} \phi(\sigma) & \text{if } \dim \phi(\sigma) = d, \\ 0 & \text{otherwise.} \end{cases}$$

One can check that $\phi_d^\#$ sends $\mathsf{Z}_d(K)$ to $\mathsf{Z}_d(L)$, and likewise for boundaries. Thus, $\phi_d^\#$ descends to a map $\mathsf{H}_d(K) \to \mathsf{H}_d(L)$ of quotient spaces, which is our homomorphism ϕ_d^*. More precisely, the following assignment of homology classes is well defined in the sense that it never sends two members of the same homology class in K to different homology classes in L:

$$\phi_d^*([x]) = [\phi_d^\#(x)].$$

From a computational perspective, one constructs a matrix representation of $\phi_d^\#$ and computes its Smith normal form in order to explicitly construct ϕ_d^*.

For a classical and theoretical account of simplicial homology, one can turn to the canonical algebraic-topology texts [28, 36]. But for a much more computational approach to homology (with cubical rather than simplicial complexes!), the reader is invited to consult [22]. There are highly optimized software libraries [29, 37, 38] for computing homology groups of various types of complexes.

4.3 Persistent Homology, Diagrams, and Stability

Suppose we start with a simplicial complex, and add a single extra vertex to it, disconnected from everything else. This change effectively increments the dimension of the zero-dimensional homology group by 1. One can easily construct examples where removing a single simplex also changes the dimensions drastically.

[6] So, there is precisely one zero in the diagonal of the Smith normal form of M_1, since K has only one connected component.

In this sense, the homology of a complex is not very stable to small changes in that complex. The antidote to this lack of stability is provided by *persistent homology*.

Persistent homology is to filtrations what homology is to simplicial complexes. Consider a filtration $\mathscr{F}$ of a simplicial complex K as shown,

$$\emptyset = \mathscr{F}_0 K \hookrightarrow \mathscr{F}_1 K \hookrightarrow \cdots \hookrightarrow \mathscr{F}_M K,$$

and note that each inclusion corresponds to a simplicial map of simplicial complexes, so one may apply the homology machine to get a sequence of Euclidean spaces connected by matrices for each dimension d:

$$\mathsf{H}_d(\mathscr{F}_1 K) \xrightarrow{\phi_d^{1\to 2}} \mathsf{H}_d(\mathscr{F}_2 K) \xrightarrow{\phi_d^{2\to 3}} \cdots \xrightarrow{\phi_d^{(M-1)\to M}} \mathsf{H}_d(\mathscr{F}_M K).$$

This structure is called a *persistence module*. The horizontal maps of homology groups are induced by chain maps arising from simplicial inclusions $\mathscr{F}_m K_d \hookrightarrow \mathscr{F}_{m+1} K_d$. Let us write $\phi^{1\to 3}$ to denote the obvious matrix product $\phi^{2\to 3} \cdot (\phi^{1\to 2})$, which gets us from the first to the third Euclidean space in our persistence module and so forth. These horizontal matrices allow one to track homological features (components, tunnels, cavities, etc.) across the entire filtration. The *p-persistent d-dimensional homology group* of the subcomplex $\mathscr{F}_m K$ is defined as the following subspace of $\mathsf{H}_d(\mathscr{F}_{m+p})$:

$$\mathsf{H}_d^p(\mathscr{F}_m K) = \phi_d^{m\to m+p}(\mathsf{H}_d(\mathscr{F}_m K)).$$

The basic idea behind this formulation is simple. Each homology class $[x]$ living in the d-dimensional homology group of $\mathscr{F}_m K$ is included into the d-dimensional homology group of $\mathscr{F}_{m+p} K$ by a string of maps on homology groups induced by simplicial inclusions. However, $\mathscr{F}_{m+p}$ contains more simplices than $\mathscr{F}_m K$ in general, so there might be a collection of $(d+1)$-dimensional simplices which fill out this cycle by making it a boundary. If this is not the case, then x has survived the journey from $\mathscr{F}_m K$ to $\mathscr{F}_{m+p} K$ safely. Otherwise, x must have met its demise at some stage q occurring before p. In the latter case, it corresponds to the trivial element [0] in the homology group of $\mathscr{F}_{m+p} K$.

In order to compute homology, we had to put matrix representations of boundary operators into Smith normal form using row and column operations with coefficients in **R**. Computing persistent homology groups requires a similar calculation, except that we now perform these operations over *polynomials* in one variable with coefficients in **R**. Using these techniques (see the canonical reference [42, Sect. 4.2] for an explicit algorithm), one can compute for each nontrivial homology class $[x]$ in $\mathsf{H}_d(\mathscr{F}_m K)$ an unambiguous interval $[b_x, d_x)$, where the *birth* $b_x \leq m$ and the *death* $d_x > m$ are defined as follows:

- b_x is the smallest ℓ such that there is some homology class $[y]$ in $\mathsf{H}_d(\mathscr{F}_\ell K)$ with $[\phi_d^{\ell\to m}(y)] = [x]$, and

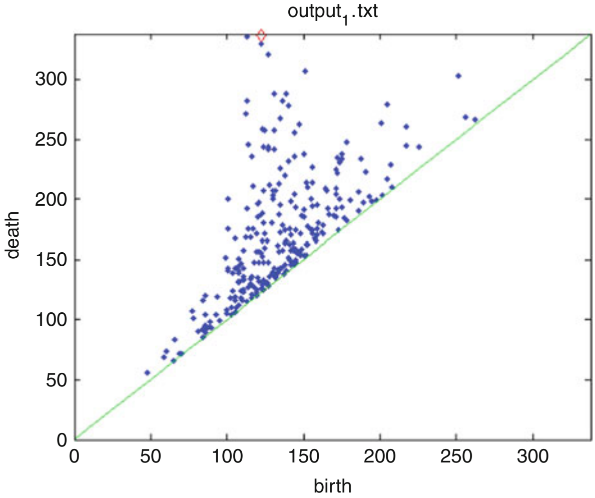

Fig. 8 A sample persistence diagram generated by the Perseus software package [29]. Births are plotted along the *horizontal axis* and deaths along the *vertical axis*. The points near the diagonal correspond to homology generators which do not persist across a large section of the filtration, and hence correspond to unstable or noisy features. On the other hand, the *dots* far from the diagonal correspond to robust features with long lifespans

- d_x is the smallest n such that $[\phi_d^{m\to n}(x)]$ is the trivial homology class $[0]$ in $\mathsf{H}_d(\mathscr{F}_n K)$.

This collection of *persistence intervals* $[b_x, d_x)$ over all such x is called the d-dimensional *persistence diagram* of the filtration $\mathscr{F}$, and it can be easily visualized as a two-dimensional cluster of points (see Fig. 8). For each x, the length $(d_x - b_x)$ measures the *lifespan* of the homology class $[x]$ across the filtration. The persistence diagram is the filtered analog of the Betti numbers in the following sense: the d-dimensional Betti number of $\mathscr{F}_m K$ is simply the number of d-dimensional persistence intervals which contain m.

Stability. There is a well-defined notion of distance between persistence diagrams, called the *bottleneck distance*. It is known [8] that the persistence diagram is stable to fluctuations in the filtration. In particular, consider a point cloud P in Euclidean space and a "noisy" version P', which is another point cloud obtained by perturbing each point of P by some distance less than a fixed $\gamma > 0$. Then, one can prove that the bottleneck distance between the dimension-d persistence diagrams of the Čech or Vietoris–Rips filtrations of P and P' is smaller than γ for every d. In this sense, the output persistence diagram is no more noisy than the input point cloud.

It is crucial to note that this stability result is a one-way street. That is, *if* P and P' are near each other, then their persistence diagrams will also be close. But it would be wrong to conclude that P and P' are close if their persistence diagrams are similar. Thus, having similar persistent homology only allows one to *conjecture* the similarity of the underlying datasets; however, having different persistent homology actually furnishes a solid *proof* that the two datasets are topologically distinct. Thus, persistent homology is better at telling things apart than at confirming their similarity.

There are various excellent resources for the persistent-homology neophyte; see [5, 6, 13, 16–18] and the references therein for many more details. The reader may also be relieved to know that using persistent homology does not require a personal desire to compute Smith normal forms of huge matrices by hand: efficient software is available for this purpose [24, 29].

5 Applications to Biological Datasets

Having established the basics of simplicial complexes and their homology, we would like to highlight some particularly appealing instances of topological inference – that is, inference based on topological techniques – from biological datasets. Selecting the right filtration to impose on a point cloud is a bit of an art form: even choosing an expedient distance function between data points requires highly specialized knowledge about the data itself, as well as a genuine understanding of the desired features which one wishes to investigate. In the absence of a general recipe that fits all possible data, the next best thing is a host of successful and interesting examples which the reader can use as signposts in his or her personal quest to build a convenient filtration.

5.1 Identification of Breast Cancer Subtypes

Breast cancer is one of the most widespread and most frequently occurring types of cancer. Since there are several variants of this cancer, considerable efforts have been made to distinguish these from each other in the search for specialized and effective treatments.

5.1.1 The Discovery of c-MYB+

A new subtype of breast cancer was detected in [31] by clustering methods acting on a filtered simplicial complex built using microarray data.

The data. A *microarray* [2] is a thin glass slide with distinguished regions – called *features* – onto which DNA molecules can attach in an orderly fashion. Using these slides, it is possible to measure efficiently as *patterns* the differences in expression between two sets of genes (from a common cell) which have been kept under different conditions. In [30], a framework called Disease-Specific Genomic

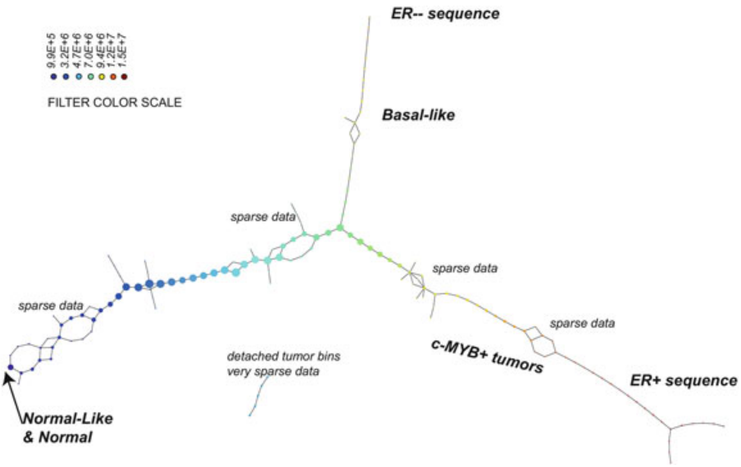

Fig. 9 Progression Analysis of Disease (PAD) results from [31] produced by Mapper [34]: the data points correspond to tumors, and their colors represent the order of magnitude of deviation from normal as measured by DSGA: *red* tumors have the largest deviation

Analysis (DSGA) was introduced, which highlights differences in expression patterns of microarrays of diseased tissue relative to a continuous range of normal phenotypes. The input data was precisely the result of DSGA performed over a sufficiently large class of normal and diseased tissues.

The complex. Nicolau et al. [30] constructed the complete simplicial complex K whose vertices T correspond to a set of tumors, and defined a function $g : T \to \mathbf{R}$ derived from the distance of each tumor from some large collection of normal phenotype tissue as yielded by regular DSGA analysis. The precise details of this distance function may be found in [31, Sect. 1.3]. Associating each simplex to the highest g-value encountered among its vertices extended g to all of K. The sublevelset filtration of g was then fed into the clustering tool Mapper [34].

The results. As shown in Fig. 9, Mapper revealed an intrinsic structure of the space of breast cancer transcriptional data that remained undetected by common clustering methods. Without any clinical or biological input except for DSGA, the construction of a suitable filtered simplicial complex followed by clustering enabled the detection of a new, unique subgroup of breast cancers called c-MYB+. These cancers are estrogen receptor-positive (ER+), and have high levels of x-MYB and low levels of innate inflammatory genes. Perhaps most importantly, there is a 100 % survival rate and no metastasis. This type of cancer does not fit into the standard classification of Luminal A/B and Normal-like subtypes of ER+ breast cancers obtained by ordinary clustering analysis.

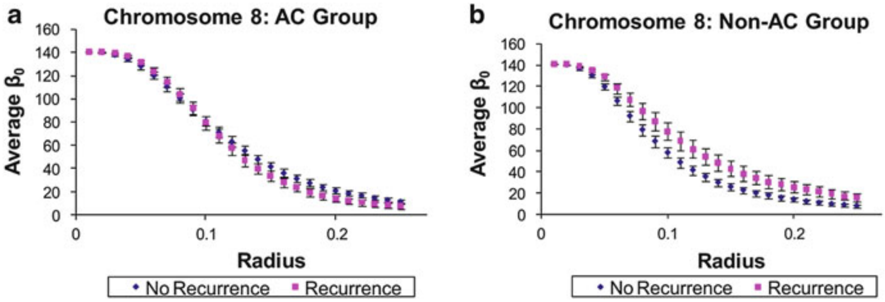

Fig. 10 Chromosome 8 plots of average Betti numbers β_0 in dimension 4 calculated for recurrent and nonrecurrent data, for radii between 0.01 and 0.25. (**a**) Patients treated with chemotherapy (AC group). (**b**) Patients not treated with chemotherapy (non-AC group). Non-AC patients have significantly higher β_0 values in the recurrent population

5.1.2 Distinguishing Between Recurrent and Nonrecurrent Subtypes

Dewoskin et al. [14] established that topological methods can partially differentiate those breast cancer subtypes which have a high recurrence rate from those which do not.

The data. *Comparative genomic hybridization* (CGH) is a method which detects chromosomal aberrations [33]. DNA from a tumor sample and from a normal reference sample are given different fluorescent labels and cohybridized onto a thin glass surface in a regular pattern. The fluorescent intensity of each region measures the differences (either *amplifications* or *deletions*) between the two samples as the logarithm of a ratio. The starting point of this analysis, therefore, is an ordered list

$$\ell = (\ell_1, \ell_2, \ldots, \ell_N)$$

of these logarithms of ratios of intensities.

The complex. One chooses an *embedding dimension* d, and creates points in $\mathbf{R}^d$ by sliding a window of width d along the list ℓ as follows. The first point is $(\ell_1, \ldots, \ell_d)$, the second one is $(\ell_2, \ldots, \ell_{d+1})$ and so forth. This creates a point cloud $P_d(\ell) \subset \mathbf{R}^d$. Although this point cloud does not retain precise knowledge of *where* the tumor DNA differs from the normal DNA, the pairwise distances are preserved and similar regions are mapped near the origin in $\mathbf{R}^d$. If the intensities are similar, then their ratio is close to 1, and hence the logarithm of the ratio is near 0. The Vietoris–Rips filtration was constructed around the point cloud $P_d(\ell)$ for various choices of dimension d.

The results. For $d = 4$, the *average zero-dimensional Betti numbers* (Fig. 10) over all of the Vietoris–Rips subcomplexes for chromosomes 8 and 11 clearly

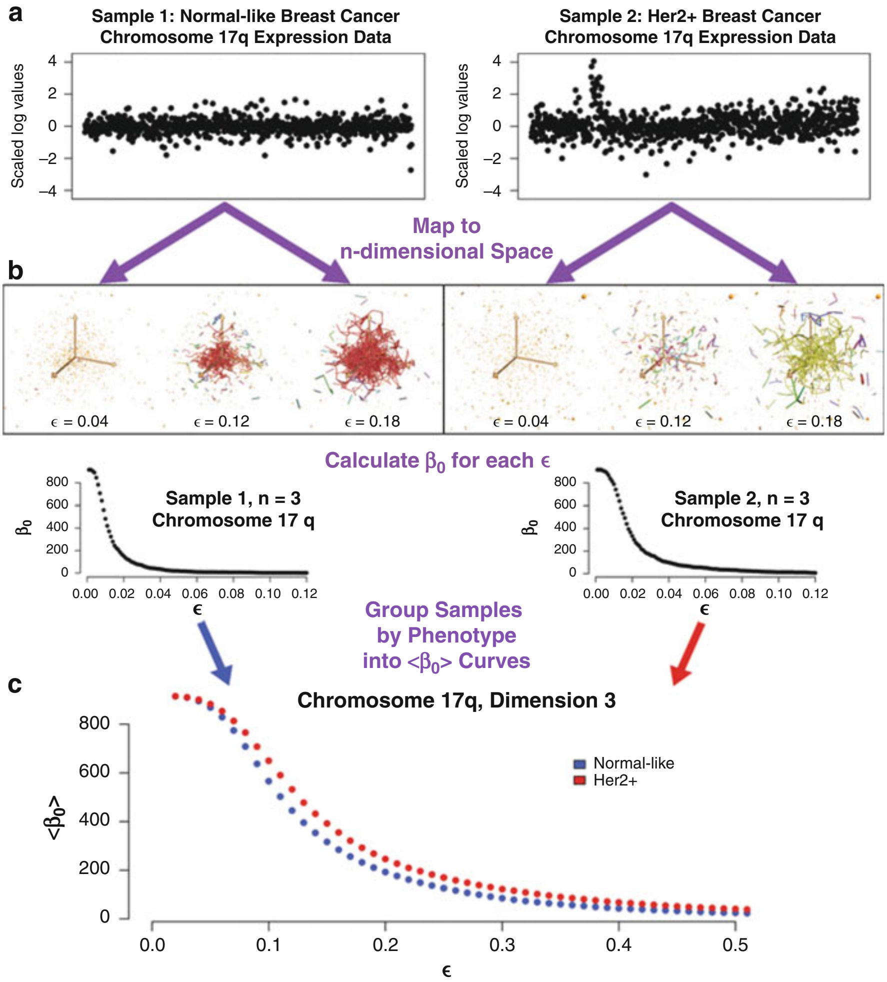

Fig. 11 Outline of the method used in [1]. (**a**) Gene expression for chromosome 17q for patients with two different types of breast cancer. (**b**) Point clouds and plots of β_0. (**c**) Plots associated with different sets of patients for a window size equal to 3. The final steps include statistical analysis for combining and correcting values

distinguished between recurrent and nonrecurrent patients who did not receive anthracycline-based chemotherapy after surgery. This method reproduced results presented in [7]. See Fig. 11 for a pictorial summary.

5.1.3 Drawing Finer Distinctions with Persistent Homology

Arsuaga et al. [1] used persistent homology instead of Betti number averages, and, starting with the same data and complex, extended the results of [14]. For an embedding dimension $d = 3$, analyzing zero-dimensional persistence diagrams had partial success in differentiating various subtypes of breast cancer. In particular, it was possible to differentiate between cancers with varying disease progression: the less aggressive types included Normal-like and Luminal A, whereas the more aggressive types were Luminal B, Basal, and Her2. The zero-dimensional persistence diagrams could differentiate intrinsic subtypes such as Basal-like and Her2 further within the class of aggressive cancers. The persistence intervals suggest that Luminal B has features in common with both the Her2 and the Basal-like subtypes.

In the future work, Arsuaga et al. hope to relate these results to cancer recurrence predictions and use the full strength of persistent homology. The fundamental question is that of which properties of breast cancers – if any – are captured by higher-dimensional homology groups. The ultimate goal is to gain insight into the periodicity of disease progression and hence select the most effective treatments.

5.2 *Analysis of Neural Structures*

A fundamental question arising from investigations of the brain's perception mechanisms is how a physical environment is mapped into the visual cortex, and how the resulting mental maps are used by the hippocampus for spatial navigation.

5.2.1 Activity Patterns in the Visual Cortex

The central thesis of [23] is that spontaneous cortical states resemble the patterns in oriented stimuli, i.e., that they have the same topology. The work of Singh et al. [35], described below, provides supporting evidence for this claim.

The data. The basic data consisted of multielectrode recordings from the primary visual cortex of a macaque[7] in two different settings: spontaneous activity when both eyes were closed, and natural image stimulation when one eye was open and exposed to a video sequence.

The complex. The recorded data was split into 10-s segments, and the five neurons with the highest firing rates were selected. The spike trains were binned into 50-ms intervals, so that each segment corresponded to 200 points. Finally, a witness

[7] *Simia inuus*, an Old World monkey.

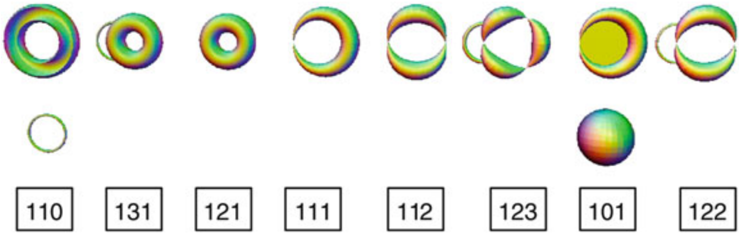

Fig. 12 Different topological signatures obtained in the experiment [35]. The *top row* contains examples of complexes, with the prescribed Betti number sequence signature $(\beta_0, \beta_1, \beta_2)$ shown below the corresponding complex

complex approximation to the Vietoris–Rips filtration was built around 35 landmark points.

The results. Singh et al. [35] computed the persistence intervals in dimensions 0, 1, and 2, at various scales in the Vietoris–Rips filtration and referred to such strings of Betti numbers as *signatures*. Several different Betti number sequences observed during the experiment are shown in Fig. 12, along with a sample complex whose homology exhibits those Betti numbers. Figure 13 shows histograms of Betti number distributions obtained for spontaneous and natural image stimulation, where the Betti numbers follow the same order as in Fig. 12. The features present for higher threshold scales correspond to more persistent features of the Vietoris–Rips filtration built around the data.

The experiment showed that the homology of a circle and sphere dominated the data, although the circle was much more prevalent during natural image stimulation than during spontaneous activity. The main difference between the two experimental settings appeared at lower thresholds, where the spontaneous activity exhibited much more diverse topological structures.

5.2.2 Activity Patterns in the Hippocampus

The hippocampus is a part of the brain which contains *place cells* – neurons that can detect location – clustered into regions called *place fields*. The hippocampus plays a central role in an animal's ability to navigate in its environment. However, the process by which visual data is converted into a spatial map in the brain remains mysterious. Dabaghian et al. [10] worked under the hypothesis that the topology of the map obtained from the place cells in the brain matches the topological features of the environment. That is, they conjectured that the brain does not have access to the geometric information and that it converts neural signals into a spatial map (similar to a subway map) of the surroundings based only on the spiking activity of the place cells [9, 19] and on the connectivity and adjacency information. They also

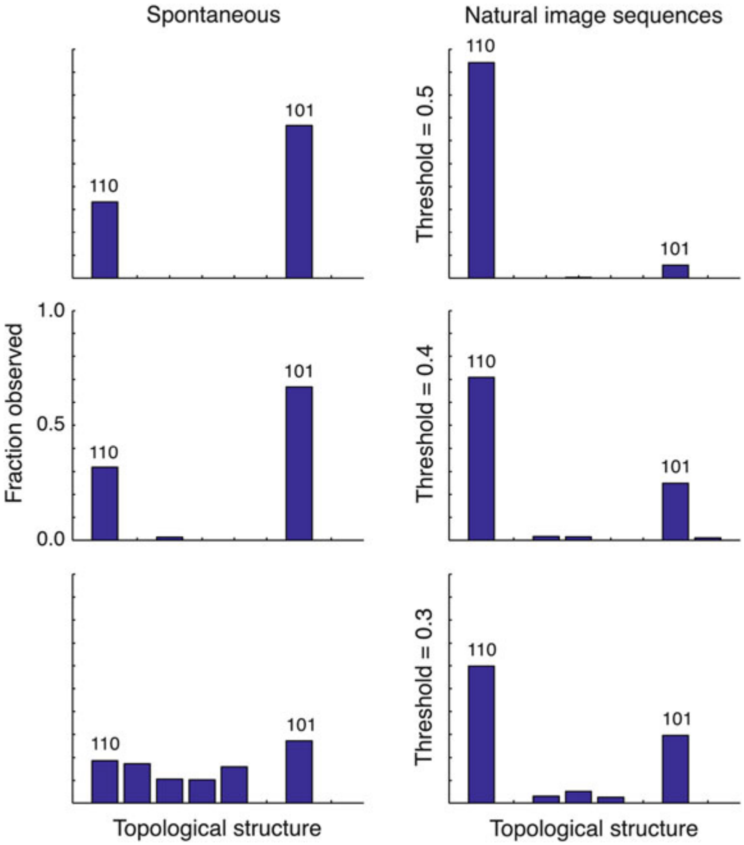

Fig. 13 Histograms of Betti number distributions obtained in the spontaneous and natural image stimulation phases of neural activity. The thresholds correspond to different Vietoris-Rips scale values. Higher thresholds correspond to more persistent features of the data. For lower threshold values, the spontaneous activity exhibits diverse topological structures, while natural image stimulation is still dominated by the homology of a circle and a sphere

assumed that the hippocampus constructs the connectivity map based on the place cell cofiring patterns.

For example, consider a rat running through a maze. As it begins to explore the environment, place fields in its hippocampus become active: in the beginning, they are be disconnected, but over time, as various navigation routes are explored, the connectivity of the active regions increases and eventually holes begin to appear.

The data. To model the activity of the place cells in a computer simulation, the authors of [10] considered the firing rate f; the size of the place field s (the part of the hippocampus that is activated when a place cell fires), of ellipsoidal shape; and the number of cells N.

The complex. A simplicial complex K was constructed as follows. Each place field was a vertex, and a d-dimensional simplex σ of K consisted of $(d - 1)$ place fields which fired simultaneously during the experiment. Let $\|\sigma\|$ denote the total number of place cells involved in the simultaneous firing. The monotone function $g : K \to \mathbf{R}$ defined by

$$g(\sigma) = 1 - c\sqrt{\frac{\|\sigma\|}{N}}$$

for any $c > 0$ provides a measure of the *dissimilarity* between the place fields which form the vertices of σ. A positive c was chosen, and sublevelsets of g were used to generate a filtration of K. This is precisely the simplicial model presented in [9].

The results. The most amazing result obtained by computing persistent homology was that, if one ignored features with very small lifespans, then the homology of K was the same as the homology of the environment. The results for different experimental conditions are summarized and explained in Fig. 14, from [10]. The top row (i) shows three different experimental configurations of the environment, but we note that (B) and (C) are topologically the same. The second row (ii) contains the mean map formation times; each dot represents a place cell with a certain (f, s, N), and the size of the dot represents the percentage of trials in which this state produced the correct outcome. The color range denotes the time needed to form the map, blue denoting a short time and red almost the whole time period. Note how the third scenario (C) contains a preponderance of blue dots, which means that it was much easier for a rat to map this configuration rather than (B), even though they are topologically indistinguishable.

5.2.3 Terminal Ganglia of Crickets

The cricket *Acheta domesticus* uses hairs on its rear appendage (called a *cercus*) to detect changes in its environment. The hairs are connected via nerve endings called *afferent terminals* to the *terminal ganglion*, one of the three dense neural centers present in the cricket's body. These hairs are broadly classified as *proximal* and *distal*, depending on their distance from the ganglion. The proximal hairs are further divided into *long, medium*, and *short* categories, whereas the distal hairs are always long. Each hair has an *orientation*, a preferred direction to which it is most sensitive.

The afferent terminals of hairs with different orientations are in different places in the terminal ganglion. Hence, the cricket's response to an external stimulus depends on the region in the terminal ganglion which is excited by the stimulus, and this region depends on the direction of the stimulus. A natural question is to determine whether there is a similar dependence for spatial stimuli: i.e., whether different spatial stimuli correspond to a spatial segregation of the terminal ganglion. This would imply that the projections of the long, medium, and short hairs in the terminal

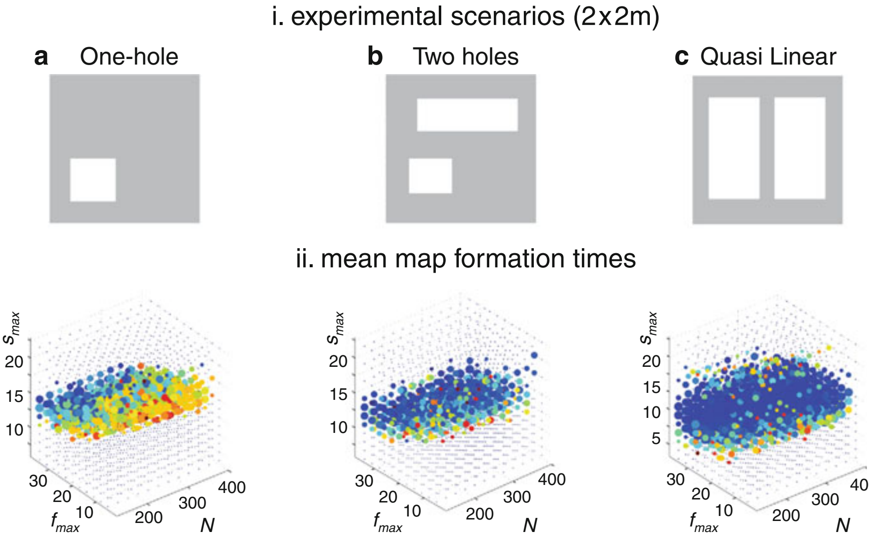

Fig. 14 (i) Three different experimental configurations of the environment: (B) and (C) are topologically identical. (ii) Point cloud approximations that reveal mean map formation times for each space configuration. Each *dot* represents a hippocampal state as defined by the three parameters (f, s, N); the size of the *dot* reflects the proportion of trials in which a given set of parameters produced the correct outcome. The color of the *dot* reflects the mean time taken over ten simulations: *blue* denotes a short time, whereas *red* stands for almost the entire period. The maximum observed time was 4.3 min for configuration (A), 11.7 min for (B), and 9.3 min for C

ganglion are concentrated in different regions of the terminal ganglion. Since the structure of afferent terminals and their attachment to the terminal ganglion is rather complicated, this question remained open until 2012. Recently, however, a positive answer was provided by Brown and Gedeon [3] using topological tools.

The data. The data came from experiments on afferent terminals [20, 21, 32], with the data points representing the three-dimensional locations of terminal endings in ganglia across a large number of crickets. This data was preprocessed via various standard methods, including Gaussian mixture models and nearest-neighbor techniques. The data for the afferent terminals of long, short, and medium hairs was isolated into three separate point clouds.

The complex. The authors of [3] constructed a Vietoris–Rips filtration around each point cloud, but used the distance $\mathbf{d}_1$. The scale-thickened version of such a complex consists of cubes rather than balls. The reason for doing this involved the large size of the dataset: cubes require less memory to store on a computer than do simplices, and there is a parallel theory of cubical homology.

The results. The authors of [3] compared the persistent homology of the initial point clouds for the long, medium, and short hairs separately, then for all pairwise unions, and finally for the complete dataset. The results are shown in Fig. 15 as

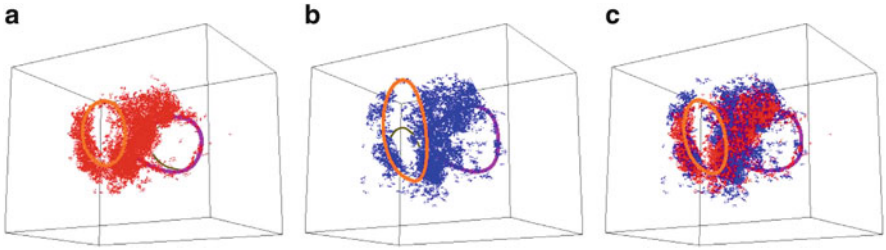

Fig. 15 Experimental data for (**a**) the short hairs, (**b**) the medium hairs, and (**c**) the medium and short hairs combined, together with the generators of the first homology. (**a**) and (**b**) have three persistent generators (*orange, purple*, and *gray*), but the last generator is filled up in the combined set

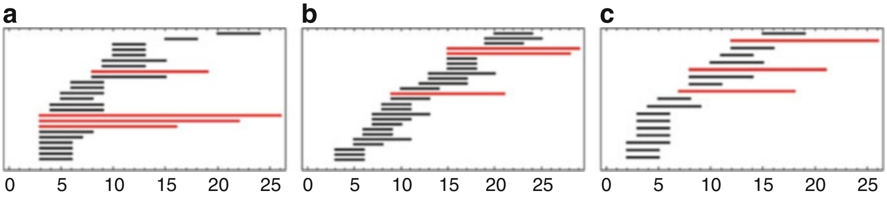

Fig. 16 Barcodes of dimension 1: β_1 persistence intervals of length more than two for the reduced datasets for the proximal hairs. (**a**) Long hairs; (**b**) medium hairs; (**c**) short hairs. Persistent generators are shown in *red*

barcodes, where each bar is drawn from the birth scale to the death scale and its length represents the lifespan of the corresponding homological feature. The persistence of the cubical filtration complex was computed using the software package cubPersistenceMD [25–27]. The persistent homologies of the individual point clouds and their unions were found to be significantly different.

As an example of the methodology, we compare the union of the short and medium proximal hairs. There are three persistent generators in each point cloud when the clouds are viewed individually. However, when one considers the combined point cloud consisting of data from both short and medium hairs, then only two of these three persistent generators remain (see Fig. 15). Computations reveal that one of the persistent generators for the medium set is filled by the terminals from the short hairs (Fig. 16).

In light of these observations, one can conclude that the nerve endings connected to the hairs are actually concentrated at different places in the terminal ganglion. Thus, there is the potential for downstream neurons to use information from the hairs. The precise nature of how these neurons synapse with the nerve endings in the terminal ganglion is unknown, and is currently under investigation.

Acknowledgements The authors thank Robert Ghrist for useful discussions, feedback and support. VN's work is supported by federal contracts FA9550-12-1-0416 and FA9550-09-1-0643. RS is grateful for the generous support by NSF 0935165 FA9550-12-1-0416 grants.

References

1. J. Arsuaga, N. Baas, D. DeWoskin, H. Mizuno, A. Pankov, C. Park, Topological analysis of gene expression arrays identifies high risk molecular subtypes in breast cancer. Applicable Algebra in Engineering, Communication and Computing. Special issue on Computer Algebra in Algebraic Topology and Its Applications. **23**, 3–15 (2012)
2. M. M. Babu, Introduction to microarray data analysis, in *Computational Genomics*, ed. by R. Grant (Taylor & Francis, 2004)
3. J. Brown, T. Gedeon, Structure of the afferent terminals in terminal ganglion of a cricket and persistent homology. PLoS ONE **7**(5), e37278 (2012)
4. G. Carlsson, Topology and data. Bull. Am. Math. Soc. (N.S.) **46**(2), 255–308 (2009)
5. G. Carlsson, V. de Silva, Zigzag persistence. Found. Comput. Math. **10**(4), 367–405 (2010)
6. G. Carlsson, V. de Silva, D. Morozov, Zigzag persistent homology and real-valued functions, in *Proceedings of the 25th Annual Symposium on Computational Geometry*, Aarhus (ACM, 2009), pp. 247–256
7. J. Climent, P. Dimitrow, J. Fridlyand, J. Palacios, R. Siebert, D.G. Albertson, J.W. Gray, D. Pincel, A. Lluch, J.A. Martinez-Climent, Deletion of chromosome 11q predicts response to anthracycline-based chemotherapy in early breast cancer. Cancer Res. **67**, 818–826 (2007). PMID: 17234794
8. D. Cohen-Steiner, H. Edelsbrunner, J. Harer, Stability of persistence diagrams. Discret. Comput. Geom. **37**(1), 103–120 (2007)
9. C. Curto, V. Itskov, Cell groups reveal structure of stimulus space. PLoS Comput. Biol. **4**, e1000205 (2008)
10. Y. Dabaghian, F. Memoli, L. Frank, G. Carlsson, A topological paradigm for hippocampal spatial map formation using persistent homology. PLoS Comput. Biol. **8**(8), e1002581 (2012)
11. S. Dantchev, I. Ivrissimtzis, Efficient construction of the Čech complex. Comput. Graph. **36**(6), 708–713 (2002)
12. V. de Silva, G. Carlsson, Topological estimation using witness complexes, in *SPBG'04 Proceedings of the First Eurographics Conference on Point-Based Graphics*, Zurich, 2004, pp. 157–166
13. V. de Silva, R. Ghrist, Coverage in sensor networks via persistent homology. Algebr. Geom. Topol. **7**, 339–358 (2007)
14. D. Dewoskin, J. Climent, I. Cruz-White, M. Vazquez, C. Park, J. Arsuaga, Applications of computational homology to the analysis of treatment response in breast cancer patients. Topol. Appl. **157**(1), 157–164 (2010)
15. H. Edelsbrunner, The union of balls and its dual shape. Discret. Comput. Geom. **13**, 415–440 (1995)
16. H. Edelsbrunner, J. Harer, *Computational Topology: An Introduction* (American Mathematical Society, Providence, 2010)
17. H. Edelsbrunner, D. Letscher, A. Zomorodian, Topological persistence and simplification. Discret. Comput. Geom. **28**, 511–533 (2002)
18. R. Ghrist, Barcodes: the persistent topology of data. Bull. Am. Math. Soc. (N.S.) **45**(1), 61–75 (2008)
19. B. Igelnik, *Computational Modeling and Simulation of Intellect: Current State and Future Perspectives*, vol. 655 (Information Science Reference, Hershey, 2011). xxix
20. G. Jacobs, F. Theunissen, Functional organization of a neural map in the cricket cercal sensory system. J. Neurosci. **16**, 769–784 (1996)
21. G. Jacobs, F. Theunissen, Extraction of sensory parameters froma neural map by primary sensory interneurons. J. Neurosci. **20**, 2934–2943 (2000)
22. T. Kaczynski, K. Mischaikow, M. Mrozek, *Computational Homology* (Springer, New York, 2004)
23. T. Kenet, D. Bibitchkov, M. Tsodyks, A. Grinvald, A. Arieli, Spontaneously emerging cortical representations of visual attributes. Nature **425**, 954–956 (2003)

24. D. Morozov, Dionysus software library, http:www.mrzv.org/software/dionysus
25. M. Mrozek, Homology software website, http://www.ii.uj.edu.pl/,mrozek/software/homology.html
26. M. Mrozek, B. Batko, Coreduction homology algorithm. Discret. Comput. Geom. **41**, 96–118 (2009)
27. M. Mrozek, T. Wanner, Coreduction homology algorithm for inclusions and persistent homology. Comput. Math. Appl. **60**(10), 2812–2833 (2010)
28. J. R. Munkres, *Elements of Algebraic Topology* (Addison-Wesley, 1984)
29. V. Nanda, Perseus: the persistent homology software, http://www.math.rutgers.edu/~vidit
30. M. Nicolau, R. Tibshirani, A. Børresen-Dale, S.S. Jeffrey, Disease-specific genomic analysis: identifying the signature of pathologic biology. Bioinformatics **23**(8), 957–965 (2007)
31. M. Nicolau, A.J. Levine, G. Carlsson, Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. PNAS **108**(17), 7265–7270 (2011)
32. S. Paydar, C. Doan, G. Jacobs, Neural mapping of direction and frequency in the cricket cercal sensory system. J. Neurosci. **19**, 1771–1781 (1999)
33. D. Pinkel, D. G. Albertson, Array comparative genomic hybridization and its applications in cancer. Nat. Genet. **37**, S11–S17 (2005)
34. G. Singh, F. Mémoli, G. Carlsson, Topological methods for the analysis of high dimensional data sets and 3D object recognition, in *Eurographics, Symposium on Point-Based Graphics*, Prague, 2007
35. G. Singh, F. Mémoli, T. Ishkhanov, G. Sapiro, G. Carlsson, D. Ringach, Topological analysis of population activity in visual cortex. J. Vis. **8**(8), article 11 (2008)
36. E. H. Spanier, *Algebraic Topology* (McGraw-Hill, New York, 1966)
37. The CAPD group, *CAPD::RedHom*, http://redhom.ii.uj.edu.pl
38. The Computational HOMology Project, *CHOMP*, http://chomp.rutgers.edu
39. The Protein Data Bank, http://www.rcsb.org
40. J. C. Venter, M.D. Adams et al., The sequence of the human genome. Science **291**(5507), 1304–1351 (2001)
41. A. Zomorodian, Fast construction of the Vietoris-Rips complex. Comput. Graph. **34**, 263–271 (2010)
42. A. Zomorodian, G. Carlsson, Computing persistent homology. Discret. Comput. Geom. **33**, 249–274 (2005)