Detecting Simulated Viral Recombination with Persistent Homology

Emil Geisler

1 Abstract

RNA viruses cause deadly and prolific diseases such as HIV and SARS-CoV-2. One of the primary challenges of treating RNA viruses is their rapid and unpredictable mutations [14]. RNA viruses are subject to random mutations over generations, but also more complex mutations broadly termed recombination events [4]. The standard method for modelling evolution in RNA viruses is phylogenetic trees [12], however phylogenetic trees are ineffective at modelling recombination events. Persistent homology is a topological data analysis method that has been found to be effective in modelling and detecting recombination events of RNA viruses [3] [5]. The goal of this project was to measure the effectiveness of persistent homology in detecting recombination through simulation of viral evolution. A simulation of viral evolution was developed which applies point mutations and recombination events to a population of RNA viruses. The resulting simulated data was analyzed using persistent homology to test the detection of the recombination events. The simulations suggested that persistent homology is variable in its detection of recombination and that certain measures of homology are preferable to others in the use of persistent homology.

2 Prior Work

In 2013, Chan, Carlsson, and Rabadàn published *Topology of Viral Evolution* [3], which introduced the use of persistent homology to model recombination in RNA viruses. While persistent homology as a tool in topological data analysis had existed prior to this study[11], Chan et al. were the first to use it to analyze RNA sequences. Since the publishing of *Topology of Viral Evolution*, there have been other studies which have built upon their work on the use of persistent homology in this setting.

The paper by Chan et al. provided a comprehensive analysis of the use of persistent homology to detect recombination. In addition to detecting recombination, they analyzed RNA viruses like the flu, where reassortment events can occur. Reassortment is another type of mutation that is not well modelled by phylogenetic trees. As a foundation to the use of persistent homology, Chan et al. proved mathematical justification that trees fail to model reassortment or recombination events. In particular, it was showed that the homology of any metric space well modelled by a tree must have trivial homology. Chan et al. introduced the idea that irreducible cycles represent one or more recombination/reassortment events, which is the main idea behind this work.

In Theresa Diaz' Thesis [5], persistent homology was applied to more RNA viruses such as MERS, SARS-CoV-2 and avian influenza. This paper was motivated and supported by Theresa's work. Camara et al. [2] modelled recombination in humans and expanded on the methodology introduced by Chan et al. Humphreys et al. [10] introduced a biologically inspired modification of persistent homology to better bring together the topological tools and biological setting.

3 Biological Background

3.1 RNA

DNA is a molecule that stores the genetic information for most organisms, including plants, animals, and fungi. Some viruses store and transfer their genetic material through DNA, but other types of viruses use RNA. DNA is composed of four *nucleotides*: guanine, thymine, adenine, and cytosine, which are denoted by G, T, A, and C respectively. RNA is also composed of four nucleotides: guanine, uracil, adenine, and cytosine, which are denoted by G, U, A, and C respectively. For this research, RNA can be thought of as a string of the letters G, U, A, C, so an example RNA sequence could be:

$$AGUACUGGCA$$

All of the genetic information of a RNA virus can be encapsulated in such a string, which vary in length. For instance, SARS-CoV-2 viruses are comprised of about $30,000$ nucleotides in their RNA sequence [17], while HIV viruses are comprised of roughly $9,500$ [8]. The goal of this project was to model the evolution of a virus given a set of RNA sequences.

3.2 Point Mutations

A common evolutionary event in viruses is called a *point mutation*, where a single nucleotide is swapped to another [6]; an example is indicated in the following expression:

$$AGUA\underline{C}UGGCA \quad \Longrightarrow \quad AGUA\underline{A}UGGCA$$

Point mutations are generally well modeled with traditional modelling techniques [3].

3.3 Recombination

Recombination is a more complex mutation that can occur within the genetic code of an RNA virus, where an entire fragment of genetic code is modified in a single mutation [15]. An example of recombination is illustrated in the following expression:

$$AG\underline{UACUG}GCA \quad \Longrightarrow \quad AG\underline{GUCAU}GC$$

In this instance, the nucleotides are replaced in the reverse order, which is called an *inversion* recombination. Recombination events are generally not well modeled or understood by traditional evolutionary modelling techniques. The overall goal of this project was to study existing models designed to detect recombination events in SARS-CoV-2.

4 Tracking Viral Evolution

4.1 Hamming Distance

Hamming distance is a metric on genetic sequences defined by the number of differences in their nucleotide strings. Hamming distance is denoted by d and illustrated in the following examples:

$$d(AC\underline{U}UGC, AC\underline{G}UGC) = 1$$

$$d(AC\underline{U}UG\underline{C}, AC\underline{G}UG\underline{A}) = 2$$

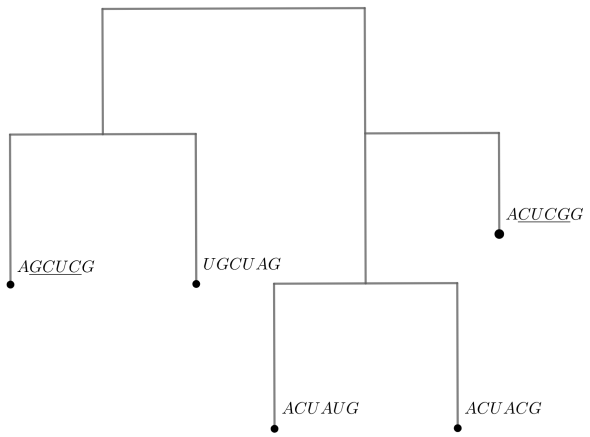


Figure 1: Failure of phylogenetic tree to model recombination.

4.2 Phylogenetic Trees

Phylogenetic trees are a standard structure to model genetic evolution [12]. There exist many algorithms to compute phylogenetic trees. In general, these algorithms place genetic sequences closer together if they are more likely to be genetically related. Because Hamming distance correlates to genetic closeness, basic phylogenitc algorithms tend to construct a tree based on Hamming distance. An example of a phylogenetic tree can be seen in Figure 1. Phylogenetic trees constructed from Hamming tend to model point mutations well because sequences that are close in Hamming distance tend to have few point mutations between them. However, recombination can result in a sequence far away in Hamming distance from the original sequence. Consider the following example where a recombination occurs between two sequences:

$$A\underline{GCUC}G, \quad UG\underline{CUAG} \quad \xRightarrow{recombination} \quad A\underline{CUAG}G$$

In this case, the sequence $A\underline{CUAG}G$ resulting from recombination is not close in Hamming distance to the sequences it originated from, so the resulting phylogenetic tree will not accurately represent the genetic mutation. This is illustrated in Figure 1, where the resulting sequence is placed on the opposite side of the tree of its parent sequences.

Even if a different algorithm for constructing the tree was used, there are still issues with representing recombination using phylogenetic trees. Recombination oftentimes occurs between more than one virus at a time [15], but trees can only represent a single parent for each sequence. In fact, from a theoretical standpoint, phylogenetic trees are provably insufficient to represent recombination events [3]. Thus, there is a need for additional evolutionary models.

5 Simplicial Homology

The evolutionary model that will be studied in this project comes from *topological data analysis* (TDA), which is an area in data analysis which seeks to analyze data with techniques from topology. The model of interest is based on the topological structures from *simplicial homology* [3]. In this section, the mathematical background necessary to understand this model will be introduced.

5.1 Simplices

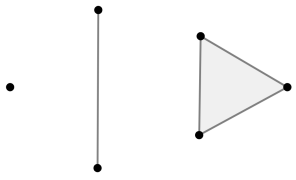


Figure 2: A 0-simplex, 1-simplex, and 2-simplex.

An *n-simplex* is a topological structure that generalizes a triangle [9]. In particular, an n-simplex is the smallest convex set in $\mathbb{R}^m$ for $m \geq n$ that contains $n+1$ distinct points $v_0, v_1, \ldots, v_n$ such that each of the vectors

$$v_1 - v_0, v_2 - v_0, \ldots, v_n - v_0$$

are all linearly independent. Thus, a 2-simplex is a triangle, a 3-simplex a tetrahedron, and so on. The points v_i are called the *vertices* of the n-simplex. An n-simplex with vertices $v_0, v_1, \ldots, v_n$ will be denoted $[v_0, v_1, \ldots, v_n]$ [9]. When computing homology the ordering of the vertices is important, so an 'n-simplex' refers to a collection of $n+1$ vertices with a specific ordering [9].

Removing a single vertex from an n-simplex $\sigma = [v_0, v_1, \ldots, v_n]$ results in an $(n-1)$-simplex denoted by $[v_0, v_1, \ldots, \hat{v_i}, \ldots, v_n]$ called a *face* of σ. The union of all faces of an n-simplex is called the *boundary* of the simplex σ and is written as $\partial\sigma$.

5.2 Simplicial Complexes

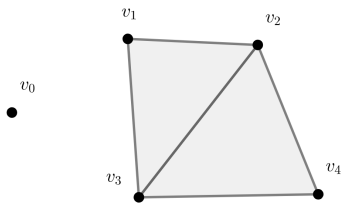


Figure 3: An example of a simplicial complex.

A *simplicial complex* K is a finite collection of simplices such that:

1. For each simplex $\sigma \in K$ and each face $\tau \subset \sigma$ of σ, τ is also a simplex of K.
2. The intersection $\sigma_1 \cap \sigma_2$ of any two simplices $\sigma_1, \sigma_2 \in K$ is another simplex in K.
3. The interior $\sigma \setminus \partial\sigma$ of each simplex $\sigma \in K$ is disjoint from every other simplex in K.

These conditions ensure that the simplices in a simplicial complex do not puncture one another or overlap in uneven sections. In particular, this structure ensures that the topology on a simplicial complex can be defined by a quotient map on its simplices. Figures 3 and 4 provide a pictorial

intuition for simplicial complexes. Figure 3 is a simplicial complex with two 2-simplices, five 1-simplices, and five 0-simplices. Notice that simplices may be disjoint.

However, the set of simplices as is the case in Figure 4 fails to be a simplicial complex because it does not satisfy criterion 3: The interior of the 1-simplex $[v_2, v_3]$ is

$$\partial[v_2, v_3] = [v_2, v_3] \setminus v_2 \cup v_3$$

Additionally, $[v_2, v_3] \cap [v_4, v_5] = v_4$. Since v_4 is distinct from the vertices v_2, v_3, we have $v_4 \in \partial[v_2, v_3]$. Therefore, $\partial[v_2, v_3] \cap [v_4, v_5] = v_4$. Therefore, the interior of $[v_2, v_3]$ is not disjoint from the simplex $[v_4, v_5]$.

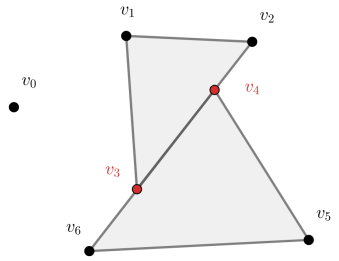


Figure 4: An example of a collection of simplices failing to be a simplicial complex.

5.3 Homology

Let K be a simplicial complex. Define $\Delta_n(K)$ to be the free abelian group with basis the n-simplices $\sigma_1, \sigma_2, \ldots, \sigma_m$ in K. This means that $\Delta_n(K)$ is a group with elements that are formal sums of the form $\sum_{i=1}^m r_i\sigma_i$ for $r_i \in \mathbb{Z}$. The group operation on $\Delta_n(K)$ is defined so that the coefficients r_i are added component wise between two elements. For example, let $x = \sum_{i=1}^m r_i\sigma_i, y = \sum_{i=1}^m s_i\sigma_i$ be two elements of $\Delta_n(K)$. Then, the abelian group operation is defined to be:

$$x + y = \sum_{i=1}^m r_i\sigma_i + \sum_{i=1}^m s_i\sigma_i = \sum_{i=1}^m (r_i + s_i)\sigma_i$$

The boundary ∂ of an n-simplex σ was previously defined to be the $(n-1)$-simplices formed by removing each vertex of σ. In terms of the free group $\Delta_n(K)$, it turns out that it is better to define the boundary by the alternating sum of its faces. Thus, let us define the *boundary maps* $\partial_n : \Delta_n(K) \to \Delta_{n-1}(K)$ for some fixed n in the following way on an n simplex $\sigma = [v_0, v_1, \ldots, v_n]$:

$$\partial_n([v_0, v_1, \ldots, v_n]) := \sum_{i=1}^n (-1)^i [v_0, v_1, \ldots, \hat{v_i}, \ldots, v_n]$$

Then, we will extend ∂_n by linearity to all elements of $\Delta_n(K)$, such that:

$$\partial_n\left(\sum_{i=1}^m r_i\sigma_i\right) = \sum_{i=1}^m r_i\partial_n(\sigma_i) \in \Delta_{n-1}(K)$$

This defines the boundary map on all elements of $\Delta_n(K)$. Additionally, since ∂_n is defined such that $\partial_n(r_i\sigma_1 + \sigma_2) = r_i\partial(\sigma_1) + \partial(\sigma_2)$, ∂_n is a group homomorphism from $\Delta_n(K)$ to $\Delta_{n-1}(K)$.

Lemma: *the composition $\partial_{n-1} \circ \partial_n : \Delta_n(K) \to \Delta_{n-2}(K)$ is the zero map.*

Proof. In the case that $n = 1$ or $n = 0$, this is trivially true since Δ_0 is the zero map by definition. Thus, assume $n \geq 2$. Take any simplex $\sigma = [v_0, v_1, \ldots, v_n] \in \Delta_n(K)$. Then, we have:

$$\partial_{n-1} \circ \partial_n([v_0, v_1, \ldots, v_n]) = \partial_{n-1}\Big(\sum_{i=1}^{n}(-1)^i[v_0, v_1, \ldots, \hat{v_i}, \ldots, v_n]\Big)$$

Applying ∂_{n-1} will remove another vertex v_j for each $j \neq i$. Let us break up the sum into cases: for $j < i$ and $j > i$. Then, we have:

$$= \sum_{j<i}\sum_{i=1}^{n}(-1)^j(-1)^i[v_0, v_1, \ldots, \hat{v_j}, \ldots, \hat{v_i}, \ldots, v_n] + \sum_{j>i}\sum_{i=1}^{n}(-1)^{j+1}(-1)^i[v_0, v_1, \ldots, \hat{v_i}, \ldots, \hat{v_j}, \ldots, v_n]$$

Notice that the signs are flipped in these two sums. In particular, any simplex for some $i < j$ in the first sum is also in the second sum as $j < i$ with the opposite sign, so the sum is equal to zero:

$$\partial_{n-1} \circ \partial_n(\sigma) = 0 \qquad \forall \sigma \in \Delta_n(K)$$

□

Another way of stating this lemma is that $\operatorname{Im}\partial_{n+1} \subseteq \operatorname{Ker}\partial_n$. A *chain complex* is a sequence of abelian groups with homomorphisms between them such that the image of one homomorphism is contained in the kernel of the next. Thus, since $\operatorname{Im}\partial_{n+1} \subseteq \operatorname{Ker}\partial_n$, the abelian groups $\Delta_n(K)$ with the maps ∂_n form a chain complex by definition:

$$\cdots \longrightarrow \Delta_{n+1}(K) \xrightarrow{\partial_{n+1}} \Delta_n(K) \xrightarrow{\partial_n} \Delta_{n-1}(K) \longrightarrow \cdots \longrightarrow \Delta_0(K) \xrightarrow{\partial_0} 0$$

Given a chain complex, we can define the nth *homology* group to be the quotient group formed by $\operatorname{Im}\partial_{n+1}$ and $\operatorname{Ker}\partial_n$. This is valid since a chain complex requires $\operatorname{Im}\partial_{n+1} \subseteq \operatorname{Ker}\partial_n$. Let H_n^Δ be the nth homology group of the chain complexes of $\Delta_n(K)$:

$$H_n^\Delta(K) = \frac{\operatorname{Ker}\partial_n}{\operatorname{Im}\partial_{n+1}}$$

The rank of $H_n^\Delta(K)$ is called the nth *Betti number* of K, and will be denoted as $\beta_n(K) = |H_n^\Delta(K)|$. Intuitively, the Betti numbers of a simplicial complex encode the number of n-dimensional holes in the complex. For instance, $\beta_0(K)$ is the number of connected components of K, while $\beta_1(K)$ is the number of one-dimensional holes. The elements of $\operatorname{Ker}\partial_n$ are called *cycles*, while the elements of $\operatorname{Im}\partial_{n+1}$

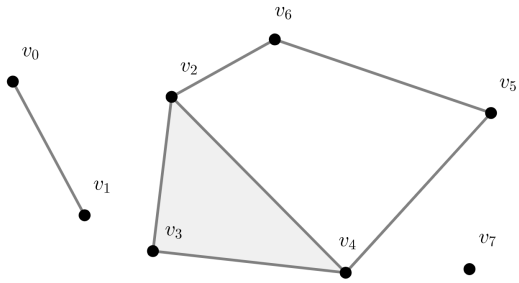
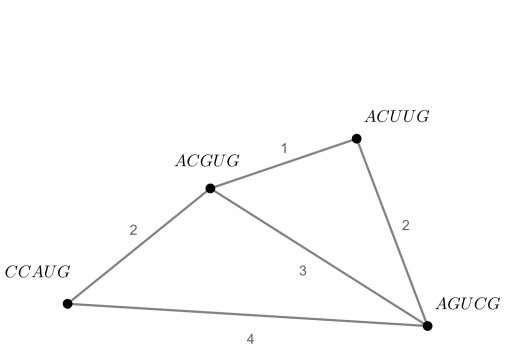
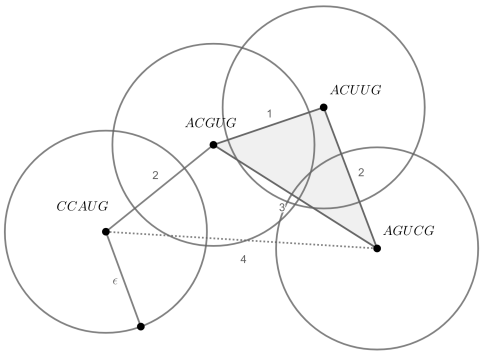


Figure 5: A simplicial complex K with $\beta_0(K) = 3$, $\beta_1(K) = 1$, and $\beta_n(K) = 0$ for $n \geq 2$



(a) RNA sequences as a point cloud.



(b) Vietoris-Rips complex.

Figure 6: Illustration of Vietoris-Rips Complexes

are called *boundaries*. The reason why the simplicial complex in Figure 5 has non-trivial homology group $H_1^\Delta(K)$ is because the cycle $[v_2, v_4] + [v_4, v_5] + [v_5, v_6] + [v_6, v_2] \in \operatorname{Ker} \partial_1$ cannot be represented as a formal sum of the boundaries of 2-simplices in K. For this reason, the holes of a simplicial complex are sometimes called *irreducible cycles*, since they are cycles that cannot be reduced to the sum of boundaries.

5.4 Vietoris-Rips Complexes

Notice that a set of RNA sequences can be thought of as a set of points in a high-dimensional metric space with distances between points defined by Hamming distance, as in Figure 6a. A *Vietoris-Rips* complex is a method to construct a simplicial complex from a point cloud as follows [11]:

1. Place a ball of radius R at each point in the point cloud for some $R > 0$, where R is the *filtration parameter*.

2. For each pair of points whose R-balls are touching, add the 1-dimensional simplex (edge) connecting them.

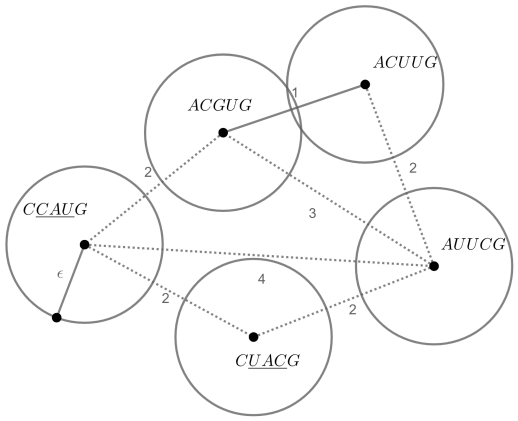
3. Add each higher dimensional simplex to the complex if each of its 1-dimensional edges are included in the complex.

5.5 Persistent Homology

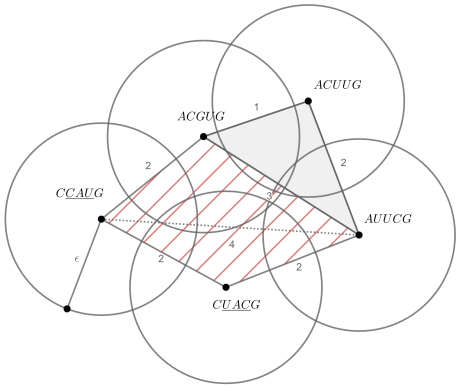
Persistent homology is a data analysis method from topological data analysis (TDA) which constructs Vietoris-Rips complexes K_R at varying values of R and then calculates the homology groups $H_n^\Delta(K)$ for each complex [5].

As the filtration parameter increases, certain elements of $H_n^\Delta(K_R)$ may persist for a wider range of filtration parameters than others. These tend to be more significant as measures of the topology of the dataset.

The use of persistent homology for evolutionary modelling is based on the hypothesis that elements of $H_n^{\Delta}(K_R)$ correspond to recombination events. The intuition behind this hypothesis is that applying point mutations will disperse an RNA sequence uniformly across Hamming space and is therefore unlikely to form holes. On the other hand, recombination events can produce large jumps in the point cloud of RNA sequences. This is more likely to produce persistent holes in the Vietoris-Rips complexes. 1-dimensional holes are hypothesized to correspond with simple recombination events between 2 or less RNA sequences, while higher dimensional holes have been hypothesized to signify more complex recombination events between multiple RNA sequences [3].



(a) Vietoris-Rips complex with $\beta_1(K) = 0$, i.e. no 1-dimensional holes.



(b) Vietoris-Rips complex with $\beta_1(K) = 1$, i.e. a single 1-dimensional hole.

Figure 7: Illustration of Persistent Homology by Vietoris-Rips Complexes

6 Previous Work

7 Methodology

7.1 Project Goal

7.2 Variables of Interest

I studied the effectiveness of persistent homology to detect recombination with respect to the following three variables of interest.

1. **Measure of Homology:**

   The persistent homology algorithm returns a list of the irreducible cycles that are in the Vietoris-Rips complexes with the filtration parameter where they first appeared and the filtration parameter where the irreducible cycle was no longer present. The *persistence* of a hole (irreducible cycle) is the difference in these filtration parameters. Therefore, there is no single number that summarizes the results of persistent homology so it is not immediately clear how to analyze numerically the results of persistent homology [3]. The term *measure of homology* in this report refers to a number that can be attributed to a single run of persistent homology to estimate the number of recombination events.

   (a) **Irreducible Cycle Rate:** The total number of holes determined at all filtration parameter values. This does not take into account the persistence of the holes, only the frequency. This measure of homology was introduced in the paper *Topology of Viral Evolution* by Chan et al. [3]

   (b) **Maximum Persistence:** The maximal persistence over all holes. This does not take into account the number of irreducible cycles, only the most significant irreducible cy-

cle. This measure of homology was introduced in the paper *Topology of Viral Evolution* by Chan et al. [3]

(c) **Sum of Persistence:** The sum of the persistence of every hole in the Vietoris-Rips complexes. A combination of the irreducible cycle rate and maximum persistence, in that the sum of persistence takes into account the number and persistence of holes. This measure of homology is novel, but inspired by the approach in [3].

2. **Type of Recombination:**

There are many types of recombination reactions that can occur within RNA viruses. Fully understanding when and why recombination events occur is a complex topic not suitable for the scope of this project. For simplicity, in this project I experimented with 4 basic types of recombination [1]. I varied the following types of recombination, illustrated in Figure 8:

(a) **Deletion:** A segment containing more than one nucleotide in the RNA sequence is deleted. In Figure 8, segment *B* is deleted.

(b) **Insertion:** A segment containing more than one nucleotide in the RNA sequence is inserted. For the sake of the simulation, the inserted RNA subsequence is randomly generated. This does not occur in nature, but randomly generated insertions are substantially simpler and behave similarly enough for the purpose of the simulations. In Figure 8, segment *B* is inserted between segments *A* and *B*.

(c) **Translocation:** Two sequences split into two fragments and exchange their genetic information by swapping fragments. For the simulation it was assumed that the pieces

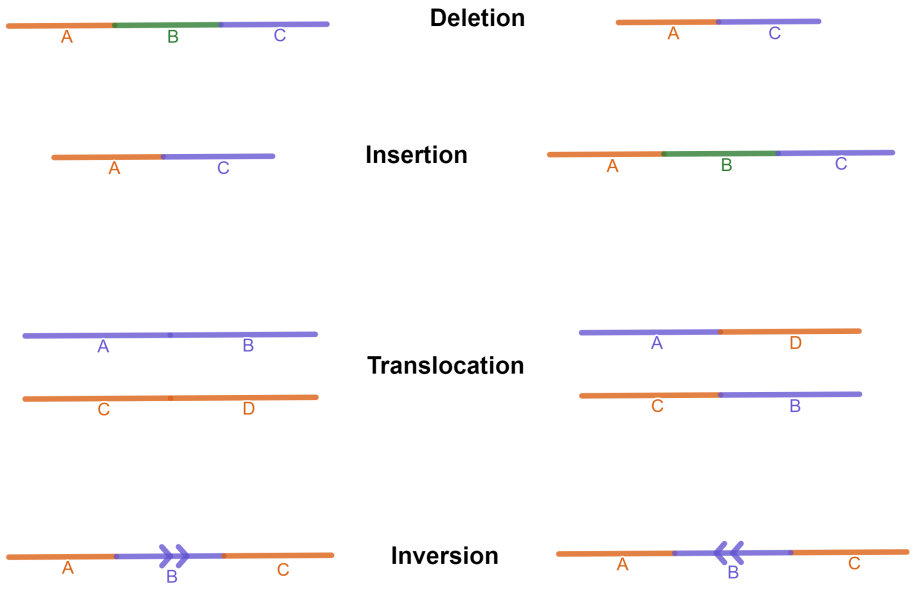


Figure 8: Types of recombination studied.

were approximately the same length. In Figure 8, the RNA sequences AB and CD exchange segments D and B respectively.

(d) **Inversion:** A segment of RNA sequence is removed, reversed, and reinserted. In Figure 8, the orientation of the segment B is emphasized to show the inversion recombination.

3. **Distance Metric:**

Recall the definition of Hamming distance. One might notice that due to insertions and deletions, different sequences may not have the same length. This would cause different sequences to be points in metric spaces of different dimension. Points in separate spaces no longer have a metric defined between them, which is necessary to construct a Vietoris-Rips complex. Thus, it is necessary to resolve the issue of different length sequences to ensure persistent homology is well defined on the given sequence space. In detecting recombination with persistent homology, I tested 3 natural choices to define Hamming distance on different length sequences. In the following examples, the symbol '$-$' denotes a deleted nucleotide in an RNA sequence.

(a) **Projection Hamming Distance:** Project both sequences to a lower dimensional space so that the standard Hamming distance can be used:

$$d(AC\underline{-}UGC, AC\underline{U}UGC) = d(ACUGC, ACUGC) = 0$$

In particular, for any set of sequences $S_1, S_2, \ldots, S_k$, all of the positions where any S_i is non-zero are not considered by the metric so that Hamming can be applied naively to the sequences with no deleted nucleotides. This is the default method to deal with deletions in the MEGA-X software [**?**].

(b) **Extended Hamming Distance:** Treat each deleted nucleotide as a separate symbol, effectively lifting the sequences to a higher dimensional space to apply standard Hamming distance:

$$d(AC\underline{-}UGC, AC\underline{U}UGC) = 1$$

This is the default method used in the BioStrings package for R from Bioconductor [13].

(c) **Proposed Distance:** Treat distance between a deleted nucleotide and a non-deleted nucleotide as having distance .5 rather than 1:

$$d(AC\underline{-}UGC, AC\underline{U}UGC) = .5$$

This distance is a valid metric. This distance may be more effective because it represents that a deleted nucleotide is not necessarily as significant as a different nucleotide. Intuitively, a sequence with some nucleotides deleted is not as different as a sequence with completely different nucleotides.

7.3 Simulation Details

The simulation proceeds as follows:

1. A random 1000 length RNA sequence is generated (the original "wild type").

2. 100 copies of the random sequence are generated to produce the starting population of RNA sequences.

3. The population of RNA sequences is simulated over 30 generations. Generations are simulated according to the Wright-Fisher model, which selects child viruses uniformly at random from the existing parent viruses from the previous generation [16]. Each child replicates the previous parent virus and then undergoes pointwise mutations and recombination events.

4. Pointwise mutations are applied at some *pointwise mutation rate*. The pointwise mutation rate is a number p with $0 \leq p \leq 1$ such that each nucleotide in the sequence is mutated with probability p. The pointwise mutation rate is varied over multiple trials. This is to measure whether persistent homology is detecting recombination events or pointwise mutations.

5. Recombination events are applied to the population of RNA sequences with respect to the *recombination rate*. The recombination rate is a number r with $0 \leq r \leq 1$ such that each sequence in the population has a recombination applied with probability r. Recombination events are applied at a random location in the sequence. Notice that 3 of the recombination types studied (deletion, insertion, inversion) must choose some recombination *length*. The algorithm chooses uniform randomly between 100 and 200 for this recombination length.

6. Pairwise distances for the population of sequences are calculated according to the distance metric of choice, e.g. projection Hamming distance, extended Hamming distance, and the proposed distance.

7. Persistent homology is computed using the ripsDiag function in the TDA package in R [7]. Only the 1-dimensional holes are computed - i.e., only the 1st homology group $H_1^\Delta(K)$. While further dimensional homology groups can detect recombination as well, the 1-dimensional holes have been found to be the most consistent in prior work [3]. Additionally, computing $H_n^\Delta(K)$ for $n > 1$ is significantly more computationally expensive.

8. The measure of homology with respect to the output of the ripsDiag function is returned.

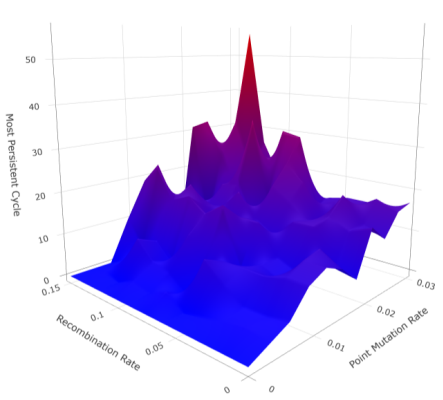
Notice that there are many constants involved in the simulation procedure. These constants were determined by testing different values to ensure persistent homology had sufficient detection of recombination. An analysis of why each constant was chosen is provided below in the following:

1. SARS-CoV-2 viruses have roughly $30,000$ nucleotides [17], so originally the sequences were chosen to have $30,000$ nucleotides. However, it was found to be too computationally expensive to simulate sequences of roughly $2,000$ Additionally, longer sequences were desirable since they are less sensitive to deletion recombination events, which were one of the recombination events modeled.

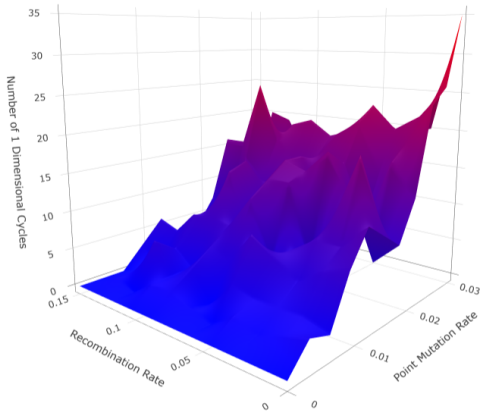
2. 100 copies of the original sequence was chosen due to computational constraints as well. More copies of the virus was found to not be computationally viable. Additionally, roughly 100 sequences tended to be used for persistent homology due to similar computation restraints [5][3].

3. Notice that increasing the number of generations by some factor p is generally equivalent to increasing the recombination and pointwise mutation rates by p. This is because recombination and pointwise mutations are simulated to occur with some probability at each generation, and so increasing the number of generations effectively increases the mutation rate. Additionally, since new sequences are chosen uniformly at randomly from the previous generation according to the Wright Fisher model, changing the number of generations does not vary the distribution of sequences on average. Therefore, the number of generations was chosen to be 30 somewhat arbitrarily while the mutation rates were kept variable.

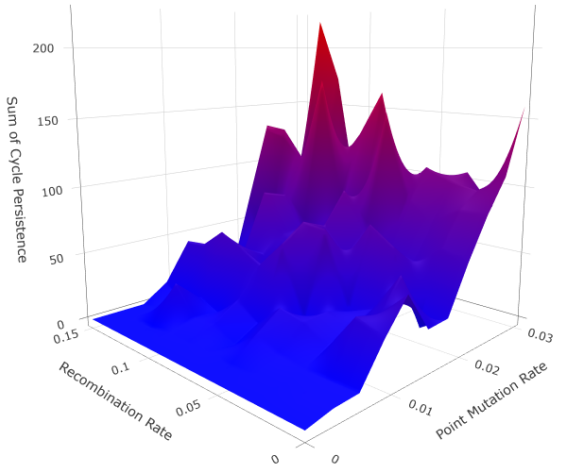
4. The length of the recombination events was chosen to be between 100 and 200 by experimentation. With too small of recombination lengths, there were no irreducible cycles found whatsoever, and so the use of persistent homology would be trivial. The simulations suggested that recombination length of around 100 led to the appearance of holes, so this range was chosen to introduce some randomness.



(a) Maximum persistence



(b) Irreducible Cycle Rate



(c) Sum of Persistence

Figure 9: Simulations varying the measure of homology with insertion recombinations.

8 Results

8.1 Maximal Persistence is Optimal for Detecting Recombination

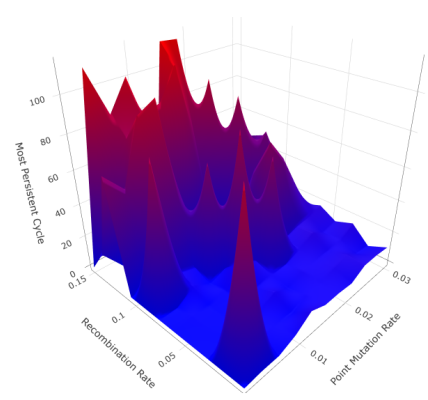
Figure 9 displays the measure of homology plotted on the y-axis from a simulation with recombination rate on the x-axis and the point-mutation rate on the z-axis. The goal of persistent homology is to detect recombination. Therefore, the most effective measure of homology is that which rises with recombination but not with the point mutation rates.

Thus, our simulations suggest that the irreducible cycle rate and sum of persistence measures of homology are ineffective at detecting recombination since the y-axis rises with the point mutation rate instead of the recombination rate. Therefore, the simulations suggest that maximum persistence is the optimal measure of homology for insertion recombinations. In experiments with other types of recombination, the results were consistent in suggesting that maximum persistence is the best indicator of recombination. Notice in particular that the irreducible cycle rate rises with

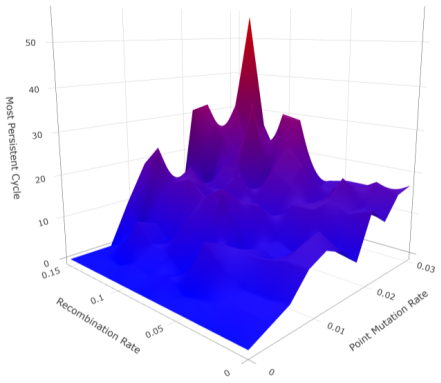
the point mutation rate especially when no recombination events are simulated. This is the part of the graph of Figure 9 panel (b) where $z = 0$, i.e. the recombination rate is 0. Additionally, there appears to be some decrease in the number of irreducible cycles as the recombination rate increases, which is the exact opposite of the desired behavior of persistent homology in this simulation. This behavior is concerning, and is not consistent with the conclusions from the paper by Chan et al. [3]

8.2 Maximum Persistence Best Detects Deletions and Inversions

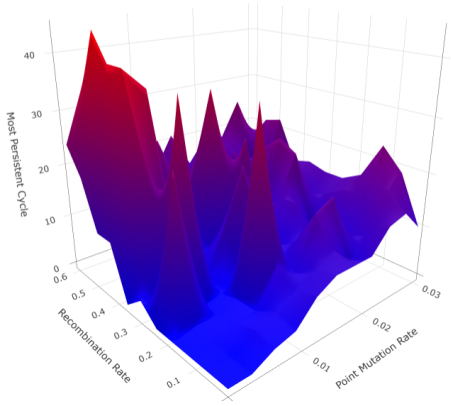
Figure 10 displays the maximum persistence of simulations with the recombination type varying between deletions, insertions, inversions, and translocations. Notice that in Figure 10 panel (a) and Figure 10 panel (c) the maximum persistence generally rises with the recombination rate - i.e., there is an upward trend along the z axis. However, the maximum persistence is not as strongly correlated in Figure 10 panel (b) and Figure 10 panel (d). In particular, in panel (d), there is a large



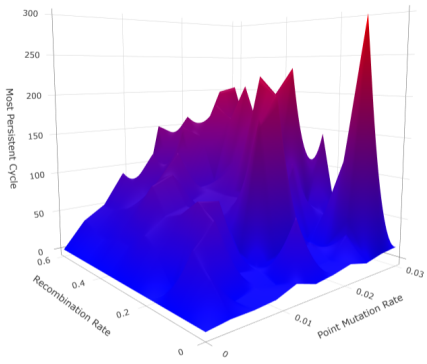
(a) Deletions



(b) Insertions



(c) Inversions

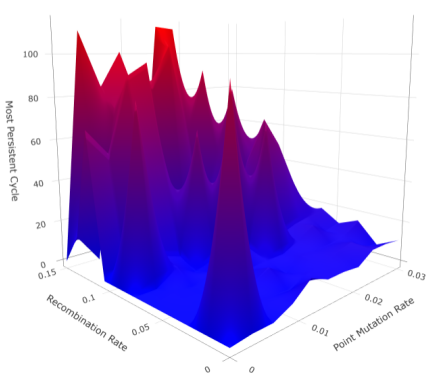


(d) Translocations

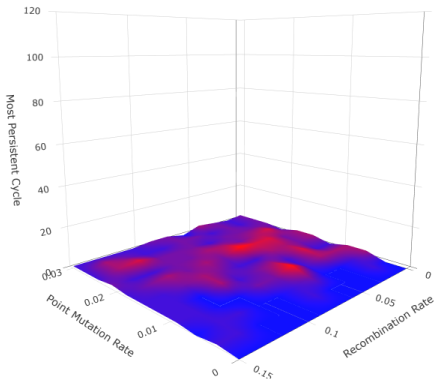
Figure 10: Simulations varying the type of recombination with y axis as the maximum persistence.

spike near the x axis with large point mutation rate but small recombination rate. This suggests that persistent homology applied to viruses with frequent translocations may estimate a large number of recombination events where there are none.

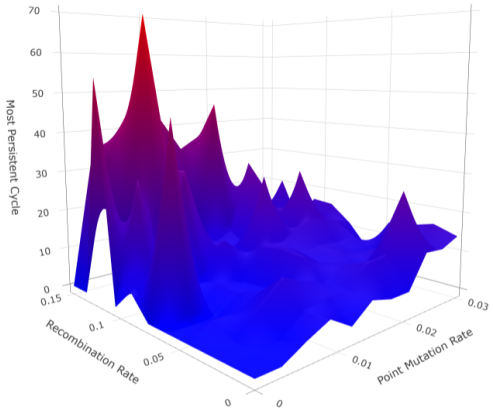
Additionally, notice that while recombination events are reasonably well detected for deletions and inversions, the results are quite "spiky". This spikiness is especially notable in panel (c) and panel (a). This spikiness suggests a significant variability in the detection of recombination. In particular, depending on the randomness of the simulation, the recombination events may be well detected or not detected at all by persistent homology. Therefore, this suggests that the detection of inversions and deletions is quite inconsistent, despite the positive correlation with recombination rate.



(a) Extended Hamming Distance



(b) Projection Hamming Distance



(c) Proposed Distance

Figure 11: Simulations varying distance metrics with deletion recombinations.

8.3 Extended Hamming Distance is Optimal for Detecting Deletions

In Figure 11, the simulation was run using different distance metrics. Additionally, the recombination type is deletion since these distances only differ on sequences with deletions. First notice that the projection Hamming distance is ineffective at detecting deletions. As more nucleotides in the sequences are deleted the sequences are projected to lower dimensional spaces and therefore more information is lost in each projection. Therefore, this distance metric is quite poor for deletions.

Both the proposed distance and the extended Hamming distance are reasonably effective at detecting recombination. However, the extended Hamming distance has a stronger correlation with the recombination rate and is therefore superior. The graph in Figure 11 panel (a) shows a significant positive trend along the z axis. However, in this experiment, since all of the recombination events were deletions, it is natural that a metric which has a higher sensitivity to deletion will fare better. To better compare the proposed distance and extended Hamming distance it would be useful to run further tests on more realistic simulations which combine deletions with other types of recombination.

9 Conclusion

9.1 Results Summary

RNA viruses cause a slew of deadly diseases, including HIV, the flu, and the recent SARS-CoV-2 pandemic. Diseases caused by RNA viruses are difficult to treat for a number of reasons - chief among them is their high mutation rate. RNA viruses can have anywhere from 100 to 10,000 times the mutation rates of DNA viruses [14]. HIV, in particular, is known for having an extremely high mutation rate, which allows the virus to escape the immune system and develop drug resistance more effectively [4].

The simulations ran suggested that persistent homology is most effective when used to analyze the maximum persistence of the holes rather than the irreducible cycle rate or sum of persistence. Therefore, a large number of irreducible cycles is less indicative of recombination than persistent cycles.

Persistent homology is most effective at detecting insertions and inversions and lease effective at detecting translocations and deletions, as supported by Figure 10. However, it is somewhat inconsistent in its detection of recombination in general, due to the varying spikes in figures 9, 10, and 11. This could be amended by combining persistent homology with another method of detecting recombination which could reduce the variance of the analysis.

The most effective distance metric on deletions is the extended Hamming distance, but further testing on different recombination types is required to make a more general statement about the efficacy of each distance metric.

One of the potential problems observed was that the recombination rate had to be quite high for persistent homology to detect it. In nature, the recombination rate is usually not as high as in these simulations, especially with respect to the point mutation rate [4]. Therefore, these simulations provide evidence that persistent homology was not as effective as desired at detecting recombination in general. However, these simulations represent a greatly simplified version of viral evolution and therefore may have excluded some key details which increase the effectiveness of persistent

homology. In particular, these simulations did not take into account the behavior of a virus, only the RNA sequence representing the genetic code of a virus. In nature, certain sequences are more viable than others and spread faster or slower than others, instead of being selected uniformly at random. These factors could drastically change the production of each generation and therefore the resulting point cloud. One might expect the point cloud to be more clustered around certain points where the viruses are most effective instead of dispersing randomly as point mutations are applied. This could lead to a point cloud characterized by small clusters of sequences spread across the sequence space.

9.2 Further Research

When using maximum persistence as a measure of homology the resulting graphs tend to be quite spiky. Additional research could prove bounds on the variability of maximum persistence with respect to simulations which would explain the spikiness of the graphs from a theoretical perspective. One might approach this problem by using the uniform randomness of point mutations to analyze the average topology of the space with only point mutations applied, and then consider how the variability of the Betti numbers of the space might change as large recombinations are applied to the sequences in the space.

Another useful way to extend the research in this project would be to drastically reduce the number of sequences and recombinations and analyze the details of why persistent homology fails or succeeds to detect recombination in specific cases. This research project employed a huge amount of trials, with 36 runs of persistent homology used per graphic (Figures 9, 10, 11), with 30 generations of 100 sequences each. While this was useful to provide a birds eye view of the effectiveness of persistent homology in detecting recombination, investigating the details of the homology of specific data sets would help to better understand how recombination affects topology.

Vietoris-Rips complexes include any simplex with dimension > 1 whenever its 1-dimensional edges are included in the complex. This is not the only way to decide whether higher dimensional simplexes are included. For instance in Euclidean space, one could include a 2-dimensional simplex if the R-balls placed around each of the triangle's three vertices have non-empty intersection. More generally, one could include an n-dimensional simplex of the R-balls around any n vertices have non-trivial intersection. This construction is called a **v**Cech Complex [11], and tends to represent the topology of point clouds in Euclidean space more faithfully than Vietoris-Rips complexes at the cost of increased computation. However, since Hamming distance is a discrete space, there is not an obvious way to construct a **v**Cech complex for RNA sequences. Therefore, it would be valuable to explore possible extensions of Hamming space to a non-discrete space to construct **v**Cech complexes for use in persistent homology.

10 Acknowledgments

Thank you so much to Javier Arsuaga for mentoring me throughout this project, Kristina Moen for working with me on the project and overcoming obstacles together, Sofia Jakovcevic for software and conceptual assistance, and the entire UC Davis Summer REU for a wonderful experience. Thank you to Theresa Diaz for the work on tackling persistent homology at Davis, and thank *you* for reading!

References

[1] Bruce Alberts. Site-specific recombination, Jan 1970.

[2] Pablo G Camara, Daniel IS Rosenbloom, Kevin J Emmett, Arnold J Levine, and Raul Rabadan. Topological data analysis generates high-resolution, genome-wide maps of human recombination. *Cell systems*, 3(1):83–94, 2016.

[3] Joseph Minhow Chan, Gunnar Carlsson, and Raul Rabadan. Topology of viral evolution. *Proceedings of the National Academy of Sciences*, 110(46):18566–18571, 2013.

[4] José M Cuevas, Ron Geller, Raquel Garijo, José López-Aldeguer, and Rafael Sanjuán. Extremely high mutation rate of hiv-1 in vivo. *PLoS biology*, 13(9):e1002251, 2015.

[5] Teresa Díaz Jordá et al. *Characterization of horizontal evolution of RNA viruses using topological data analysis*. PhD thesis, Universitat Politècnica de València, 2020.

[6] BD Editors. Point mutation: Definition, types, examples — biology dictionary, Nov 2016.

[7] Brittany T. Fasy, Jisu Kim, Fabrizio Lecci, Clement Maria, David L. Millman, and Vincent Rouvreau. *TDA: Statistical Tools for Topological Data Analysis*, 2021. R package version 1.7.7.

[8] Subgroup 'Assessment of Pathogens Transmissible by Blood' German Advisory Committee Blood (Arbeitskreis Blut). Human immunodeficiency virus (hiv). *Transfusion medicine and hemotherapy : offizielles Organ der Deutschen Gesellschaft fur Transfusionsmedizin und Immunhamatologie*, 2016.

[9] Allen Hatcher. *Algebraic topology*. Cambridge University Press, 2005.

[10] Devon P Humphreys, Melissa R McGuirl, Miriam Miyagi, and Andrew J Blumberg. Fast Estimation of Recombination Rates Using Topological Data Analysis. *Genetics*, 211(4):1191–1204, 02 2019.

[11] Jisu Kim, Jaehyeok Shin, Frédéric Chazal, Alessandro Rinaldo, and Larry Wasserman. Homotopy reconstruction via the cech complex and the vietoris-rips complex. *arXiv preprint arXiv:1903.06955*, 2019.

[12] Geetika et al. Munjal. Phylogenetics algorithms and applications. *Ambient Communications and Computer Systems: RACCCS-2018*, 2018.

[13] H. Pagès, P. Aboyoun, R. Gentleman, and S. DebRoy. *Biostrings: Efficient manipulation of biological strings*, 2021. R package version 2.60.1.

[14] Kayla M Peck and Adam S Lauring. Complexities of viral mutation rates. *Journal of virology*, 92(14):e01031–17, 2018.

[15] Etienne Simon-Loriere and Edward C. Holmes. Why do rna viruses recombine? *Nature Reviews Microbiology*, 9(8):617–626, Jul 2011.

[16] Paula Tataru, Maria Simonsen, Thomas Bataillon, and Asger Hobolth. Statistical Inference in the Wright–Fisher Model Using Allele Frequency Data. *Systematic Biology*, 66(1):e30–e46, 08 2016.

[17] Fan Wu, Su Zhao, Bin Yu, Yan-Mei Chen, Wen Wang, Zhi-Gang Song, Yi Hu, Zhao-Wu Tao, Jun-Hua Tian, Yuan-Yuan Pei, et al. A new coronavirus associated with human respiratory disease in china. *Nature*, 579(7798):265–269, 2020.