Quantifying Reticulation in Phylogenetic Complexes Using Homology

Kevin Emmett
Department of Physics
Department of Systems Biology
Columbia University
New York, New York 10027
kje2109@columbia.edu

Raul Rabadan
Department of Biomedical Informatics
Department of Systems Biology
Columbia University
New York, New York 10032
rr2579@cumc.columbia.edu

ABSTRACT

Reticulate evolutionary processes result in phylogenetic histories that cannot be modeled using a tree topology. Here, we apply methods from topological data analysis to molecular sequence data with reticulations. Using a simple example, we demonstrate the correspondence between nontrivial higher homology and reticulate evolution. We discuss the sensitivity of the standard filtration and show cases where reticulate evolution can fail to be detected. We introduce an extension of the standard framework and define the median complex as a construction to recover signal of the frequency and scale of reticulate evolution by inferring and imputing putative ancestral states. Finally, we apply our methods to two datasets from phylogenetics. Our work expands on earlier ideas of using topology to extract important evolutionary features from genomic data.

Categories and Subject Descriptors

J.3 [**Life and Medical Sciences**]: Biology and genetics; G.2.2 [**Graph Theory**]: Hypergraphs

General Terms

Theory

Keywords

topological data analysis, reticulate evolution

1. INTRODUCTION

Evolutionary relationships are often depicted using trees. From a topological perspective, trees have a simple structure, being contractible to a point. However, several evolutionary processes involve the exchange of genetic material by mechanisms which cannot be modeled by tree. These processes are collectively referred to as *reticulate evolution*, examples of which include species hybridization, bacterial gene transfer, and homologous recombination. As molecular sequence data accumulates, the importance of these processes has become increasingly apparent [6]. Here we expand on the use of ideas from topological data analysis, primarily persistent homology, to characterize reticulate evolution.

Persistent homology computes topological invariants from point cloud data [3]. The application of persistent homology to molecular sequence data was introduced in [4], where recombination rates in viral populations were estimated by computing L_p norms on barcode diagrams. In that paper, it was shown that persistent homology provides an intuitive quantification of reticulate evolution in molecular sequence data by measuring deviations from tree-like additivity. While that approach has proved successful at capturing large scale patterns of reticulate evolution, the sensitivity for detecting specific reticulate events is lower. This decreased sensitivity can be due to either incomplete sampling or weakly supported reticulations. Here, we introduce an approach for imputing latent ancestors into the data that increases the quantitative signal detected by persistent homology. Our approach is built on the *median graph* construction. Median graphs form the basis for a large number of phylogenetic network algorithms and are closely related to split decompositions of finite metrics [2, 1]. A common desire is an approach to quantify the complexity of the resulting construction. We show that using persistent homology of the median closure set is a fast and efficient way to characterize the phylogenetic incompatibility in the dataset.

The structure of the paper is as follows. In Section 2 we review the application of persistent homology to sequence data. We present two simple examples in which the standard filtration fails to capture reticulation. In Section 3 we introduce the median closure as an extended construction on the original vertex set. We show how the persistent homology of this construction recovers quantitative signal of phylogenetic incompatibilitiy. Finally, in Section 4 we present examples of our approach on two real sequence datasets.

2. PERSISTENT HOMOLOGY FOR SEQUENCE DATA

In this section we briefly review the ideas in [4] as they relate to the application of persistent homology to sequence data. Throughout, we assume biallelic data under an infinite sites model with no back mutation.

2.1 Persistent Homology

BICT 2015, December 03-05, New York City, United States

DOI 10.4108/eai.3-12-2015.2262454

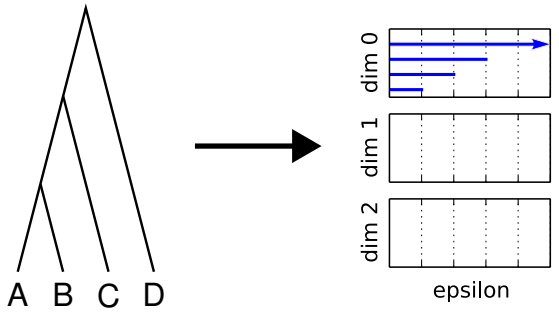

Figure 1: A tree topology is contractible and will have vanishing higher homology, as reflected in the barcode diagram.

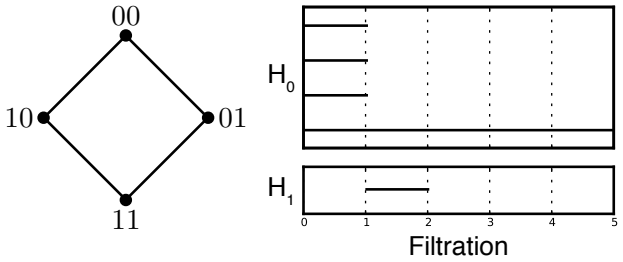

Figure 2: The fundamental unit of reticulation. (A) The four gametes represent an evolutionary loop. (B) The barcode diagram with nonvanishing H_1 in the interval $[1, 2)$.

Persistent homology computes topological invariants representing information about the connectivity and holes in a dataset. A dataset, $S = (s_1, \ldots, s_N)$, is represented as a point cloud in an l-dimensional space, where l is the length of the sequences. From the point cloud, a nested family of simplicial complexes, or a filtration, is constructed, parameterized by a filtration value ϵ, which controls the simplices present in the complex. The standard filtration is Vietoris-Rips, in which a a simplex is present at scale ϵ if the pairwise distance between each element in σ is less than ϵ. The filtration is represented as a list of simplices defined on the vertices of S, annotated with the ϵ at which the simplex appears. Given a filtration, the persistence algorithm is used to compute homology groups. The 0-dimensional homology (H_0) represents a hierarchical clustering of the data. Higher dimensional homology groups represent loops, holes, and higher dimensional voids in the data. Each feature is annotated with an interval, representing the ϵ at which the feature appears and the ϵ at which the feature contracts in the filtration. These filtration values are the *birth* and *death* times, respectively. The topological invariants in the filtration are represented in a barcode diagram, a set of line segments ordered by filtration value on the horizontal axis.

2.2 Evolution

In the standard model of evolution, novel genotypes arise via mutation during reproduction. In this case, evolutionary relationships will be accurately modeled as a bifurcating tree. A tree is trivially contractible, and hence has vanishing higher homology (see Figure 1). This result was proven for sequence data in [4]. What was not shown was the inverse statement, that vanishing higher homology implies tree-like evolution.

A simple test for the presence of reticulation is given by the *four gamete test*. The test states that the simultaneous presence of haplotype patterns 00, 01, 10, and 11 is incompatible with strict vertical evolution. Failing the four gamete test provides direct evidence for reticulate evolution. One way to quantify recombination in a set of sequences is the Hudson-Kaplan test, which counts the minimum number partitions required in the data such that within each partition all sites are compatible [8]. However, the Hudson-Kaplan test gives no further information about evolutionary relationships.

The four gametes can be considered the fundamental unit of recombination. Topologically, this unit represents a loop, as shown in Figure 2. Persistent homology identifies nonvanishing H_1 homology in the interval $[1, 2)$. We can give an interpretation to each vertex: there is a common ancestor, two parents, and a recombinant offspring. In general, we do not *a priori* know which sequences played which role in a given loop, effectively the same as the problem of rooting a phylogenetic tree. Persistent homology is then simply a method for counting the number of such loops in the data, across all genetic scales.

In considering small examples of this form we often encountered situations in which the four gamete test indicated reticulate evolution, but persistent homology failed to detect a loop, as discussed in the two examples below.

Example 1. Consider the sequences $s_1 = 000$, $s_2 = 100$, $s_3 = 010$, and $s_4 = 111$. The four-gamete test identifies incompatibility between sites 1 and 2. However, persistent homology of the four sequences does not capture this reticulation. To understand why, consider s_1 to be the common ancestor, s_2 and s_3 to be parents, and s_4 to be a descendant of a reticulate event. In this scenario, we can infer that there was an ancestral recombinant sequence, $s_r = 110$, which was not sampled. The failure to find a loop is due to the ancestral and parent sequences collapsing before connecting with the recombinant offspring, as shown in Figure 3A. In general, for a loop to be detected, the two internal distances must be greater than any of the four external distances. In this case, the internal distance from parent 1 (s_2) to parent 2 (s_3), d_{23} is equal to the distances from each parent to the sampled descendent of the recombinant (d_{24} and d_{34}). This is an example of incomplete sampling lowering the detection sensitivity, even in cases where incompatible sites are present.

Example 2. This example is taken from [10]. Consider the sequences: $s_1 = 0000$, $s_2 = 1100$, $s_3 = 0011$, $s_4 = 1010$, and $s_5 = 1111$. The four-gamete test identifies incompatibilities between sites 1 and 3, 1 and 4, 2 and 3, and 2 and 4. Performing the Hudson-Kaplan test yields a partition between sites 2 and 3, however [10] show a minimum of two reticulate events are required to explain the data. Using the standard filtration, the complex contracts completely at $\epsilon = 2$, and no higher homology will be detected. In this case, the two reticulations interact in such a way that s_3 now sits equidis-

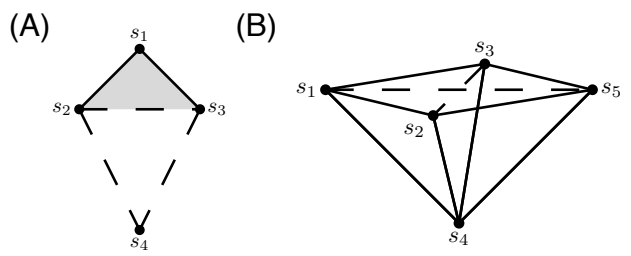

Figure 3: Two examples in which the standard filtration fails to identify reticulate evolution. (A) In this example, the ancestral sequences collapse before forming a loop with the recombinant offspring. (B) In this example, multiple recombinations interact to create a degeneracy, and the entire complex collapses immediately. (From Song and Hein [10])

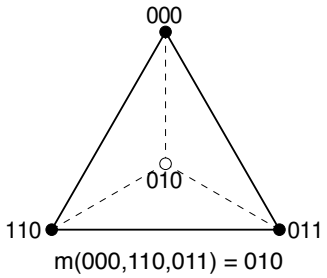

Figure 4: The median is defined as the majority allele at each position. The median closure imputes the median into the original vertex set.

tant from the other four sequences. Had s_3 not been in the data, we would have had an example very similar to Example 1, with the interpretation of one recombination event. In this example we observe that multiple reticulate events can interact in complicated ways, obscuring the signal from persistent homology.

3. THE MEDIAN COMPLEX

The median complex is an alternative construction on sequence data aimed at recovering signal of phylogenetic incompatibility using homology. First, we define the median of a set of aligned sequences.

Definition 1. For any three aligned sequences a, b, and c, the *median* sequence $m(a, b, c)$ is defined such that each position of the median is the majority consensus of the three sequences.

Consider the example shown in Figure 4. Here we have the three sequences $a = 000$, $b = 110$, and $c = 011$. Taking the majority allele at each position, the median is $m = 010$.

Next, we define the *median closure*. Given an alignment S, the median closure, $\bar{S}$, is defined as the vertex set generated from the original set S that is closed under the median operation,

$$\bar{S} = \{v \colon v = m(a, b, c) \in \bar{S} \,\forall\, a, b, c \in \bar{S}\} \qquad (1)$$

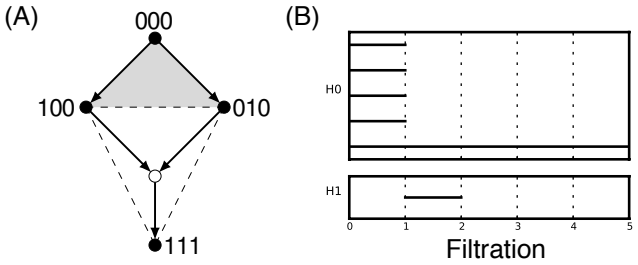

Figure 5: One median node (white node), which acts as the recombinant offspring of s_2 and s_3. One H_1 loop detected in the interval $[1, 2)$.

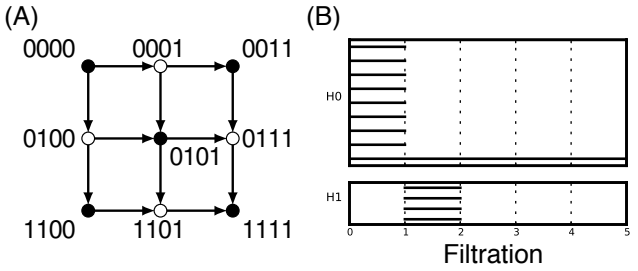

Figure 6: Four median vertices (white nodes). Four H_1 loops now detected in the interval $[1, 2)$.

We obtain the median closure $\bar{S}$ by repeatedly applying the median operation to all sequence triplets until no new sequences are added. The median closure consists of the original vertex set augmented by the computed medians. We informally refer to topological complexes formed from the median closure as *median complexes*. We can then compute persistent homology on the new vertex set.

Filtrations on median graphs have been defined previously [5], but not using explicit sequence representations. To the best of our knowledge, quantification of the complexity of these objects has not been measured using homology. We now revisit our two examples from Section 2.

Example 1. One median vertex, $m(s_2, s_3, s_4) = 110$, as shown in Figure 5. This vertex, labeled s_r, acts as the recombinant offspring of s_2 and s_3. Persistent homology now detects an H_1 loop in the range $\epsilon = [1, 2)$ formed between s_1, s_2, s_3, and s_r. s_4 is interpreted the descendant of s_r.

Example 2. Four median vertices, as shown in Figure 6. Persistent homology now detects four H_1 intervals in the range $\epsilon = [1, 2)$. In this case, the median closure now overestimates the minimum number of recombinations required. This example shows a potentially complicating aspect of the median closure in that specific H_1 features are no longer identifiable with specific reticulate events.

4. EXAMPLES

Here we consider two standard datasets from the phylogenetics literature. In both examples, the standard filtration yielded no higher homology. We generated the median clo-

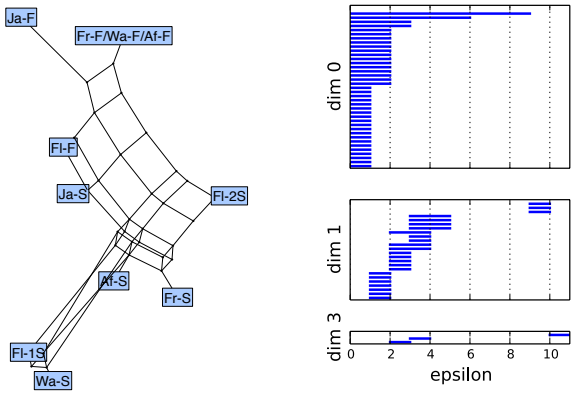

Figure 7: Recombination in *D. melanogaster*. Persistent homology identifies several complex reticulations in the population.

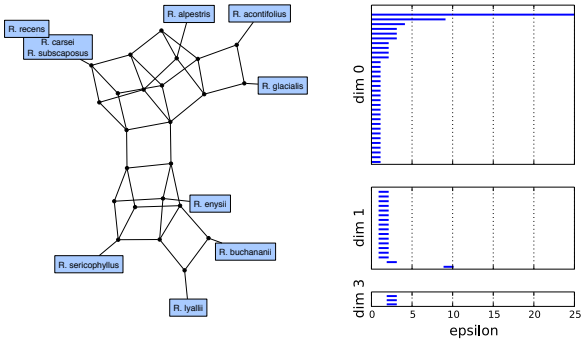

Figure 8: Species hybrdiziation in genus *Ranunculus*. Persistent homology identifies two populations separated by complex reticulations.

sure and computed homology on that. Datasets are represented using a triangle-free network construction, which approximates the computed homology.

4.1 *D. melanogaster* Data

A benchmark dataset in studying recombination is the Kreitman data [9]. The dataset consists of eleven sequences (nine unique) of the Adh locus from *Drosophilia melanogaster* collected from various geographic locations, with 43 segregating sites. The Hudson-Kreitman test yields 6 reticulate events. Computing the median closure expands the dataset to 46 vertices. Here we have non-trivial homology: 32 H_1 loops and 3 H_3 loops. In the visualized network, the complex reticulations (H_3) are localized to the bottom-most samples. The H_1 reticulations, on the other hand, are not very localized and persist across geographic regions. The barcode plot is shown in Figure 7.

4.2 *Ranunculus* Data

Natural hybridization occurs frequently in plants. Here we examine reticulation in the maturase K (matK) protein in nine species from genus *Ranunculus*. This data is originally from [7]. From nine initial species, the median closure has 32 vertices. Persistent homology is computed and the barcode diagram shown in Figure 8. Looking at H_0, we identify two clusters of species. Further, we identify 17 H_1 loops and 3 H_3 loops. Comparing with the *D. melangaster* data, reticulation at this locus is both smaller in scale (shorter bars at small filtration values) and less frequent (fewer total bars). Additionally, the complex reticulations are localized within each H_0 cluster.

5. CONCLUSIONS

Persistent homology can capture and quantify complex patterns of reticulation in genomic data. The standard Vietoris-Rips filtration is susceptible to reduced sensitivity due to incomplete sampling or interactions between reticulations. Constructing the median closure of the original sequence set increases the topological signal of reticulation. Future work will focus on efficient implementations of constructing this closure.

6. ACKNOWLEDGMENTS

KE and RR are supported by NIH grants U54-CA193313-01 and R01-GM117591-01. KE thanks Daniel Rosenbloom for useful discussions.

7. REFERENCES

[1] H.-J. Bandelt and A. W. Dress. A canonical decomposition theory for metrics on a finite set. *Advances in Mathematics*, 92(1):47–105, 1992.

[2] H.-J. Bandelt, P. Forster, and A. Röhl. Median-joining networks for inferring intraspecific phylogenies. *Molecular Biology and Evolution*, 16(1):37–48, 1999.

[3] G. Carlsson. Topological pattern recognition for point cloud data. *Acta Numerica*, 23:289–368, May 2014.

[4] J. Chan, G. Carlsson, and R. Rabadan. Topology of Viral Evolution. *Proceedings of the National Academy of Sciences*, 110(46):18566–18571, Nov. 2013.

[5] A. Dress, K. Huber, and V. Moulton. Some variations on a theme by Buneman. *Annals of Combinatorics*, 1(1):339–352, Dec. 1997.

[6] J. P. Gogarten and J. P. Townsend. Horizontal gene transfer, genome innovation and evolution. *Nature*, 3(9):679–687, Aug. 2005.

[7] K. T. Huber, V. Moulton, P. Lockhart, and A. Dress. Pruned median networks: a technique for reducing the complexity of median networks. *Molecular Phylogenetics and Evolution*, 19(2):302–310, 2001.

[8] R. R. Hudson and N. L. Kaplan. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics*, 111(1):147–164, 1985.

[9] M. Kreitman. Nucleotide polymorphism at the alcohol dehydrogenase locus of Drosophila melanogaster. *Nature*, 304(5925):412–417, Aug. 1983.

[10] Y. S. Song and J. Hein. Constructing minimal ancestral recombination graphs. *Journal of Computational Biology*, 12(2):147–169, 2005.