Topological Data Analysis and Computational Biology

Discussion with the article ”Inference of ancestral recombination graphs through topological data analysis”

Hyelim Lee

Université Côte d’Azur, Biot, 06410, France
hye-lim.lee@etu.univ-cotedazur.fr

Keywords: Topological Data Analysis · Ancestral recombination graphs · Phylogenetics · Computational Biology

1 Summary

1.1 Problems

The evolutionary process involves coordinating key themes in biology. Phylogenetics seeks to reconstruct the evolutionary history through the comparison of genomes of related organisms. The lack of a well-established universal framework capable of capturing evolutionary relationships beyond the trees can make identification and quantification of evolutionary processes difficult.[1] While several such frameworks have been proposed, existing frameworks such as phylogenetic networks are limited in their ability for biological interpretation also it is unclear how these inconsistencies arose historically, what genomic regions were involved, or how frequently an exchange happened. Ancestral recombination graphs (ARGs) provide potential explanations of the observed data in terms of a progression of recombination and mutation events. Nevertheless, the computational difficulty of creating ARGs explaining hundreds of sequences restricts these uses. Moreover, The construction of minimal ARGs, containing the minimum number of recombination events required to explain the sample in absence of convergent evolution and back-mutation, is an NP-hard problem[2]. The existing frameworks can be summarized as Table 1.

The authors aim to overcome this by employing Topological Data Analysis (TDA) methods to infer Ancestral Recombination Graphs (ARGs). TDA extends topology to finite metric spaces, inferring topological features from sampled data points. It provides robust characterizations by ensuring small data fluctuations result in minor changes in inferred features. TDA’s attribute lies in summarizing topological features as sets of intervals called barcodes, which encode the frequency and scale of events like recombination. TDA is important because it can capture the scale of inferred topological characteristics, which are represented by

Table 1. Bioinformatics tools for genetic recombination in genetic data.

Method	Characteristics	Advantages	Limitations
Tree-like representations	Represent evolutionary relationships as hierarchical trees	Simple to interpret and visualize; widely used in phylogenetics	Limited in capturing complex evolutionary processes such as recombination and reassortment
Phylogenetic networks	Extend trees to graphs, allowing for loops and reticulations	Capture inconsistencies with trees; provide large-scale representations	Limited biological interpretation of loops and historical inconsistencies
Ancestral recombination graphs (ARGs)	Explicitly accommodate recombination and mutation events	Useful in association mapping, SNP genotyping or inference of the frequency and scale of recombination	computational infeasibility of constructing ARGs for hundreds of sequences (NP-hard problem)

barcodes that indicate ranges of scales where particular features, like loops that indicate recombination events, are present in the inferred space. However, these traditional TDA implementations have limitations because they are not tailored to the biological problem they are intended to solve. In particular, it only use information about genetic distances between sequences, thus discarding the entire structure of character separations and missing the numerous recombination events needed to describe the data.

1.2 Contributions

The authors introduce a novel application of TDA to genomic data, providing a theoretical framework for capturing the genetic scale and frequency of recombination events. They describes three methods aimed at analyzing phylogenetic relationships and their implications for the analysis.

Topological ARGs It aims to reduce the complexity of constructing minimal ARGs by introducing a class of ultra-minimal ARGs that are easier to analyze. tARGs simplify an NP-hard problem into a more tractable topological problem, allowing for the analysis of large samples of sequences without the need to explicitly construct the ARGs. They offer a more parsimonious account of recombination events by focusing on those that involve genetically close parental sequences.

Persistent Homology and Recombination Inference It aims to infer information about recombination events by analyzing the loops in tARGs using persistent homology. This method allows for the detection of recombination events and provides valuable information about the genetic distances between recombining sequences. It uses barcodes to represent the persistence of topological features, giving insight to understand the size and lifespan of loops in the tARGs which correspond to recombination events.

The Barcode Ensemble of a Sample The barcode ensemble can be interpreted as counting and quantifying the scale of recombination events in a variation of Ancestral Recombination Graphs (ARGs). The barcode ensemble incorporates information about the full structure of characters in the sample, largely increasing the sensitivity of persistent homology to recombination and providing information on the location of the recombination breakpoints in the sequence. Also, It provides additional phylogenetic information like bounds on mutational distances between recombining sequences, offering a balance between fast analysis and rich phylogenetic detail.

In this approach, they show that by systematically sampling subsets of segregating sites and performing TDA, it is able to identify most of the necessary recombination events identified by bound methods, providing a significant improvement of past methods in terms of interpretation and sensitivity. Moreover, tARGs they proposed can be not only considered explicit, concise, and interpretable phylogenetic representations like minimal ARGs, but also has the advantage that this phylogenetic information can be obtained in polynomial time, which allows processing of hundreds of sequences over minimal ARGs.

1.3 Biological Discovery

The biological discovery achieved in this article lies in the ability to interpret and quantify different evolutionary processes, particularly recombination events, by analyzing large numbers of genomes. The authors present five examples to illustrate the application and interpretation of barcode ensembles in molecular phylogenetics.

A simple example This example involves a sample of four genetic sequences with seven binary characters. It demonstrates how the information in the barcode ensemble directly maps to features of ultra-minimal Ancestral Recombination Graphs (ARGs), providing insights into the genetic scales associated with recombination events.

Genetic exchange between two divergent populations This example showcases the use of persistent homology in large datasets consisting of several hundred sequences from two sexually reproducing populations exchanging genetic material at a low rate. It highlights the ability to distinguish among various biological settings with similar recombination rates using phylogenetic information contained in the barcode ensemble.

Human leukocyte antigen (HLA) locus This example focuses on 250 and 100 kilobase regions in the HLA and MS32 loci of approximately 100 individuals, where several meiotic recombination hotspots localize. It shows a large degree of consistency between the recombination rates and the genomic position of recombination events detected by the barcode ensembles. The distribution of mutational distances associated to recombination events is qualitatively consistent with coalescent arguments.

Human MS32 mini-satellite locus This example involves a sample of 97 individuals from the LWK population, sequenced by the 1,000 Genomes Project Consortium. As in the previous example, the genomic position of the recombination events detected by the barcode ensemble was consistent with the recombination rate across this region, as determined by the African-American recombination map.

Darwin's finches This example consists of the genetic sequences of 112 Darwin's finches, belonging to 15 different species inhabiting the Galápagos archipielago and Cocos Island. The first-homology barcode ensemble contains 13 recombination events, mostly involving samples from multiple species and usually including samples from the genus Certhidea, the most ancestral lineage among the genera present in the sample. These results add support to the evidence for genetic introgression. Their analysis also reveals that the crossover breakpoints of these events localize at four different genomic regions within the 9 megabase scaffold that they have considered in this example.

These examples collectively showcase the utility of barcode ensembles in capturing and interpreting evolutionary histories, recombination events, and genetic relationships across various biological contexts.

2 Discuss

The strengths and weaknesses of the methods presented in this paper are as follows.

2.1 The Strengths

Computational Effectiveness The persistent homology has a polynomial time complexity, making it computationally efficient and scalable to large datasets. It allows for the analysis of large datasets and the inference of complex evolutionary relationships. In terms of running time, most of the five samples take about 10 20 minutes to calculate, so computational effectiveness can be confirmed.

Robustness It is robust to small fluctuations in the data, providing reliable characterizations of the data and capturing the genetic scale and frequency of recombination events.

Interpretable TDA methods provide interpretable summaries of recombination events and genetic scales, enabling the identification and quantification of evolutionary processes such as migration and recombination.

2.2 The Weaknesses

Curse of Dimensionality The sensitivity of persistent homology to detect topological features decreases with the sparseness of the data and the dimensionality of the phase space.[5] In this paper, the authors approach to the problem with building a barcode ensemble, but the solution is often not unique, so another approach can be adapted to deal with the dimensional problem.

Inability to process alignments with gaps The challenge with processing alignments with gaps using tARGs arises from the nature of topological methods, which rely on the continuity and proximity of data points to infer topological features.[3] Gaps in alignments can arise from multiple sources, including historical insertion or deletion events.[4] Sometimes, alignment gaps exist simply because data corresponding to certain sequences have not been acquired. These gaps disrupt the continuity that is necessary for tARGs to accurately infer the underlying topological structure of the data.

3 Suggest Potential Improvements

To address the limitation of dimensional reduction with invariance, suitable algorithms so-called Mapper for dimensional reduction can be utilized. On the other hand, when computed on a point cloud X sampled from an underlying object X such as a manifold, the Mapper is actually an approximation of a limiting object, called the Reeb space. So, It is important to prove convergence of mapper to the limiting Reeb space. The extended persistence diagrams which are a set of topological descriptors that capture the homology of the Mapper complex at different scales can be used as a powerful tool to prove the convergence of the Mapper algorithm to the limiting Reeb space, providing a formal way to assess the topological features of Mappers. [7]

To address the limitation of Topological ARG in processing alignments with gaps, one potential improvement could be to preprocess the alignment data to remove or fill in gaps before applying the Topological ARG method. This could be done implementing algorithms to impute gaps in genetic sequences can be a solution such as statistical models or machine learning techniques to predict the most likely SNPs in the gap positions based on the surrounding sequence information. Additionally, it may be useful to compare the results obtained using Topological ARG with those obtained using other methods that are capable of handling gapped alignments, in order to ensure that the results are consistent and accurate.

References

1. W. F. Doolittle, “Phylogenetic Classification and the Universal Tree,” Science, vol. 284, no. 5423, pp. 2124–2128, Jun. 1999, doi: 10.1126/science.284.5423.2124.
2. L. Wang, K. Zhang, and L. Zhang, “Perfect Phylogenetic Networks with Recombination,” Journal of Computational Biology, vol. 8, no. 1, pp. 69–78, Feb. 2001, doi: 10.1089/106652701300099119.
3. F. Chazal and B. Michel, “An introduction to Topological Data Analysis: fundamental and practical aspects for data scientists.” arXiv, Feb. 25, 2021. doi: 10.48550/arXiv.1710.04019.
4. T.-K. Seo, B. D. Redelings, and J. L. Thorne, “Correlations between alignment gaps and nucleotide substitution or amino acid replacement,” Proceedings of the National Academy of Sciences, vol. 119, no. 34, p. e2204435119, Aug. 2022, doi: 10.1073/pnas.2204435119.
5. P. G. Cámara, “Topological methods for genomics: present and future directions,” Curr Opin Syst Biol, vol. 1, pp. 95–101, Feb. 2017, doi: 10.1016/j.coisb.2016.12.007.
6. A. E. Shikov, Y. V. Malovichko, A. A. Nizhnikov, and K. S. Antonets, “Current Methods for Recombination Detection in Bacteria,” International Journal of Molecular Sciences, vol. 23, no. 11, Art. no. 11, Jan. 2022, doi: 10.3390/ijms23116257.
7. M. Carrière and R. Rabadán, “Topological Data Analysis of Single-Cell Hi-C Contact Maps,” in Topological Data Analysis, N. A. Baas, G. E. Carlsson, G. Quick, M. Szymik, and M. Thaule, Eds., in Abel Symposia. Cham: Springer International Publishing, 2020, pp. 147–162. doi: 10.1007/978-3-030-43408-3_6.