Detecting Simulated Viral Recombination with Topological Data Analysis

Emil Geisler

UC Davis Summer REU

August 11, 2021

Motivation

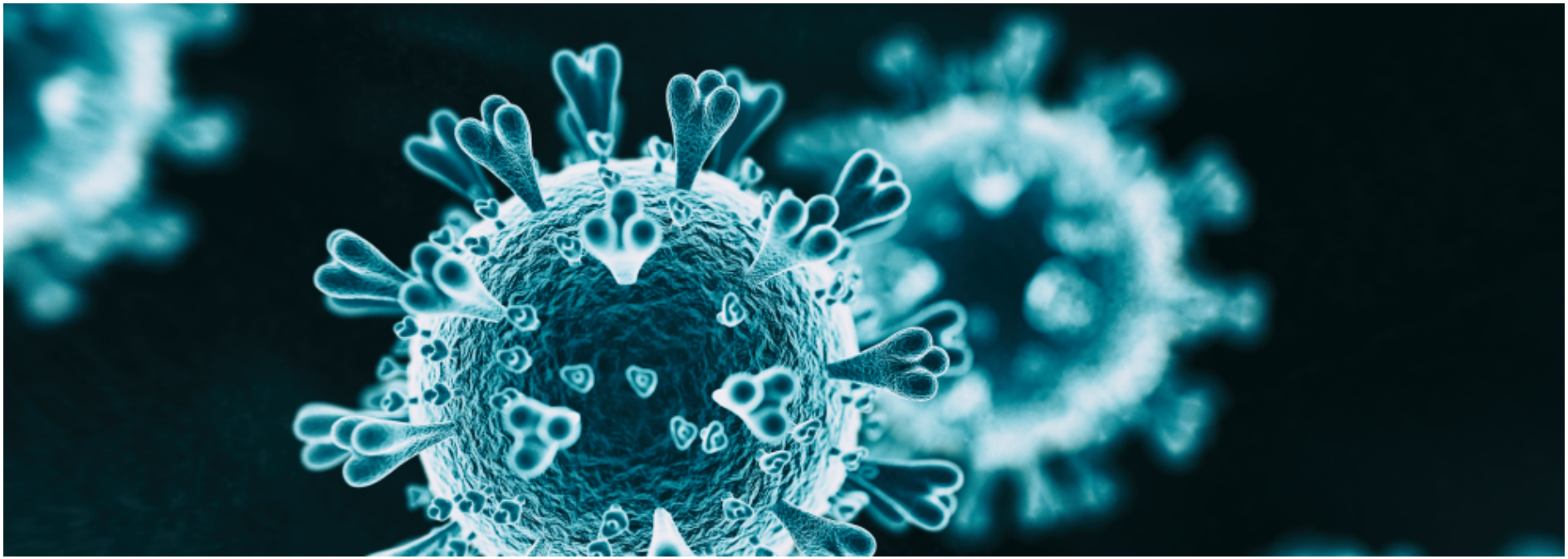


Figure: Image of HIV from the CDC

Review of RNA

RNA is a molecule which stores the genetic code for RNA viruses, such as HIV and SARS-CoV-2. RNA is comprised of 4 nucleotides labelled A, C, U, G:

Example nucleotide sequence: $ACUUCGUAUCG\ldots$

Review of RNA

RNA is a molecule which stores the genetic code for RNA viruses, such as HIV and SARS-CoV-2. RNA is comprised of 4 nucleotides labelled A, C, U, G:

Example nucleotide sequence: $ACUUCGUAUCG\ldots$

Question: Given a set of viral nucleotide sequences, can we determine the evolution of a virus?

Pointwise Mutations and Recombination

Pointwise Mutation:

$$AC\underline{U}UCGUGC \Rightarrow AC\underline{G}UCGUGC$$

Pointwise Mutations and Recombination

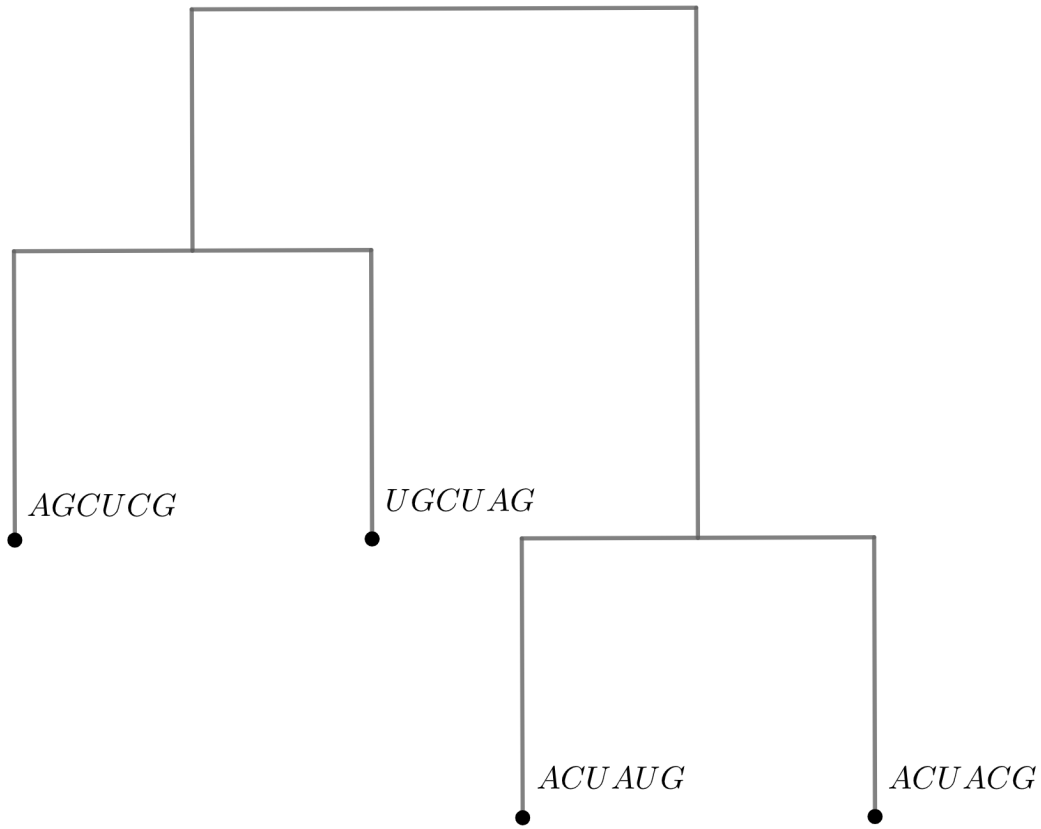
Pointwise Mutation:

$$AC\underline{U}UCGUGC \Rightarrow AC\underline{G}UCGUGC$$

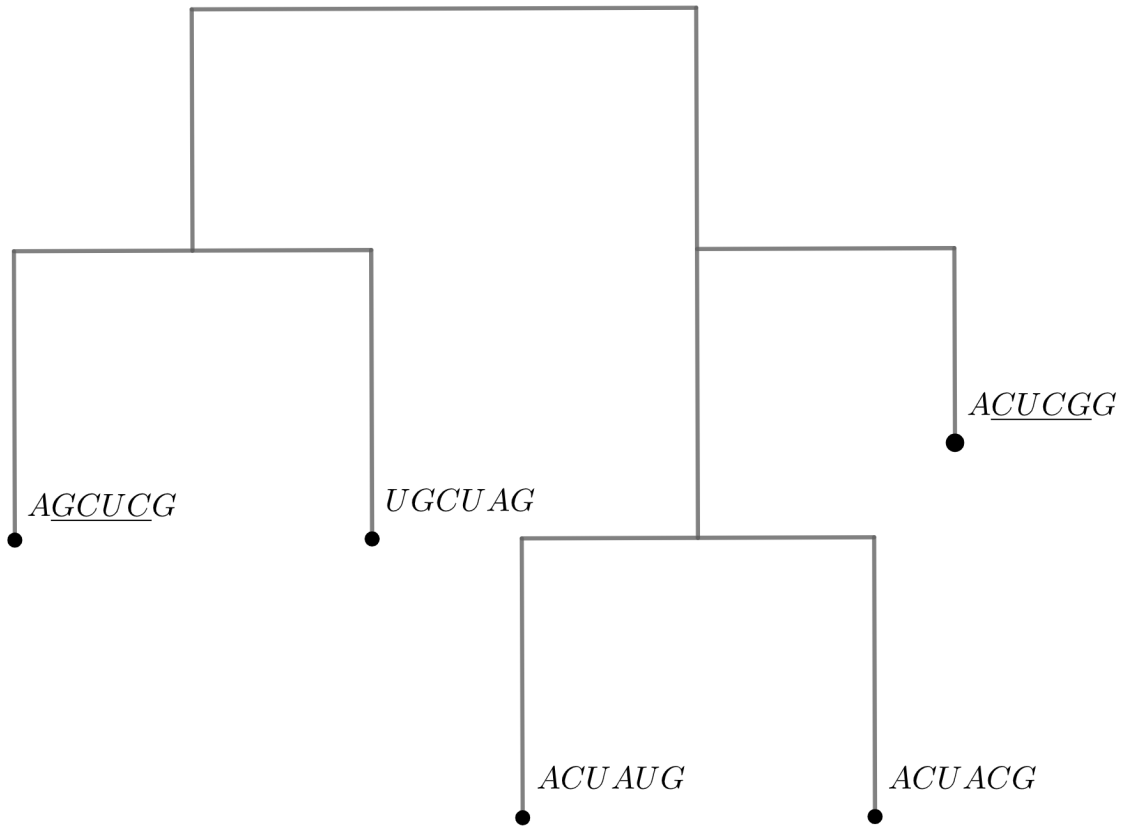
Recombination:

$$AC\underline{UUCGUG}C \Rightarrow AC\underline{GUGCUU}C$$

Phylogenetic Trees



Phylogenetic Trees on Recombination



Hamming Distance

The *Hamming Distance* between two sequences is the number of their nucleotide differences:

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

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

Hamming Distance

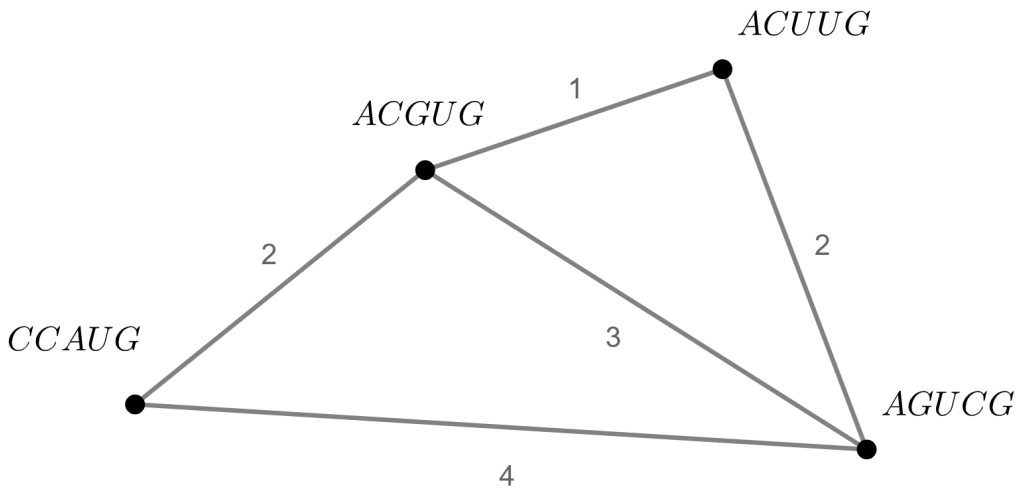
The *Hamming Distance* between two sequences is the number of their nucleotide differences:

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

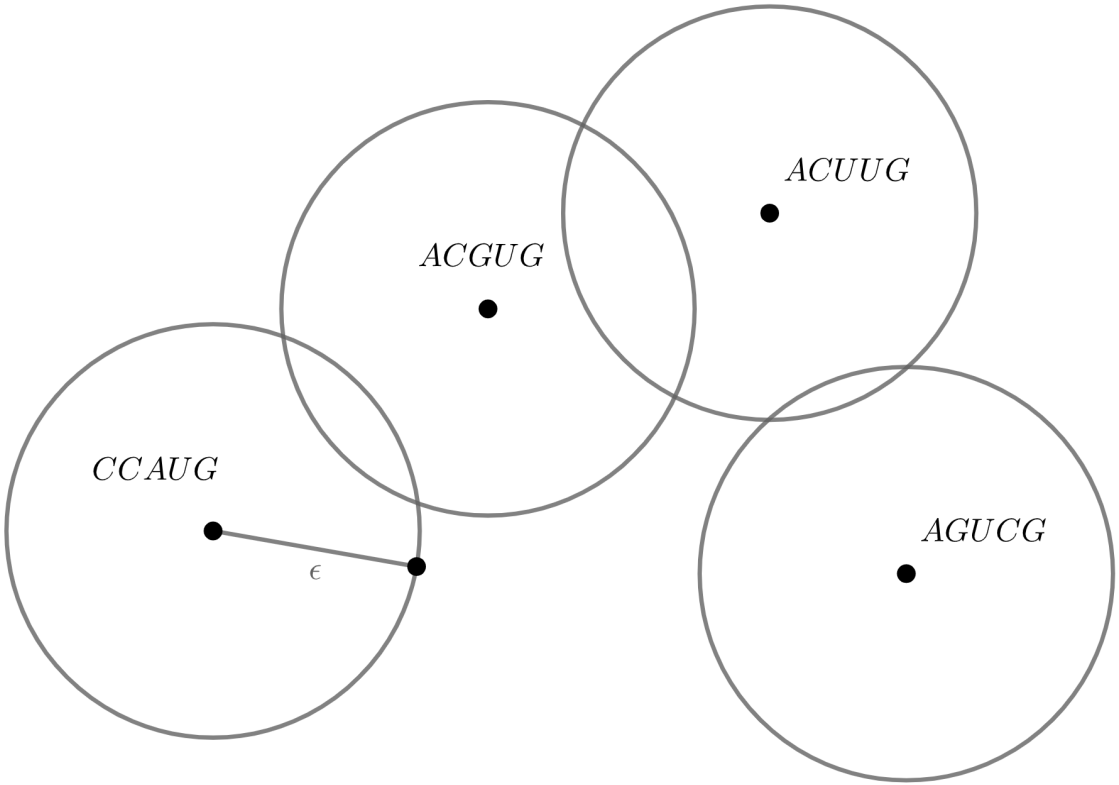
$$d(AC\underline{U}UG\underline{C}, AC\underline{G}UG\underline{A}) = 2$$

Hamming distance allows us to treat nucleotide sequences of length n as points in an n dimensional metric space.

Nucleotide Sequences as a Point Cloud



Topology of Point Cloud



Simplicial Complexes

A *simplex* is an

n dimensional generalization of a triangle.

A *simplicial*

complex is a collection of simplices.

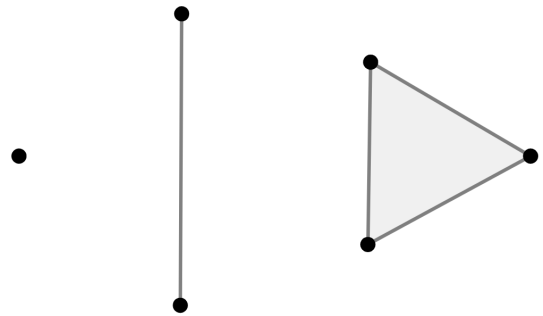
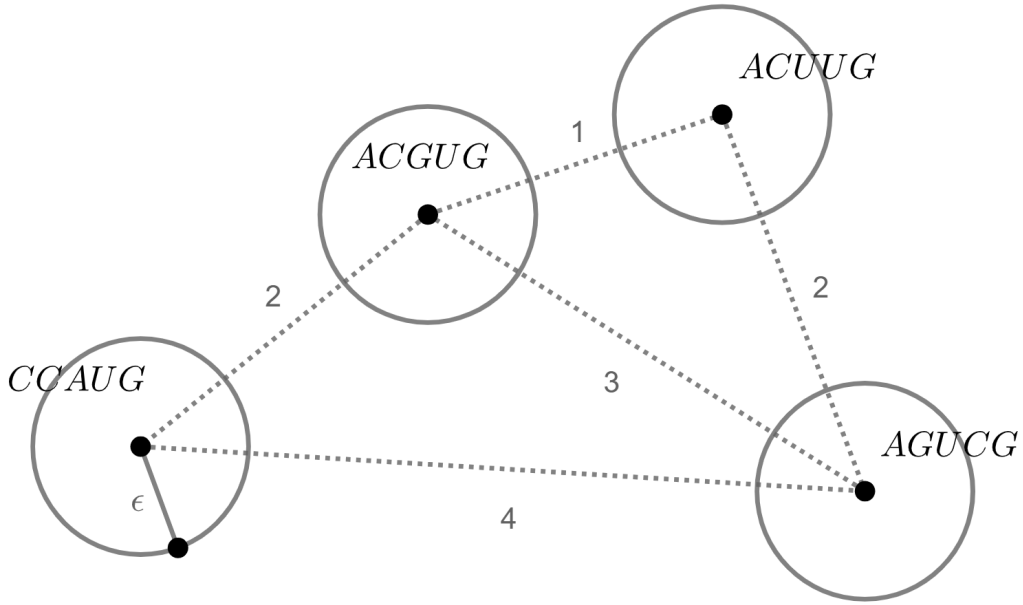
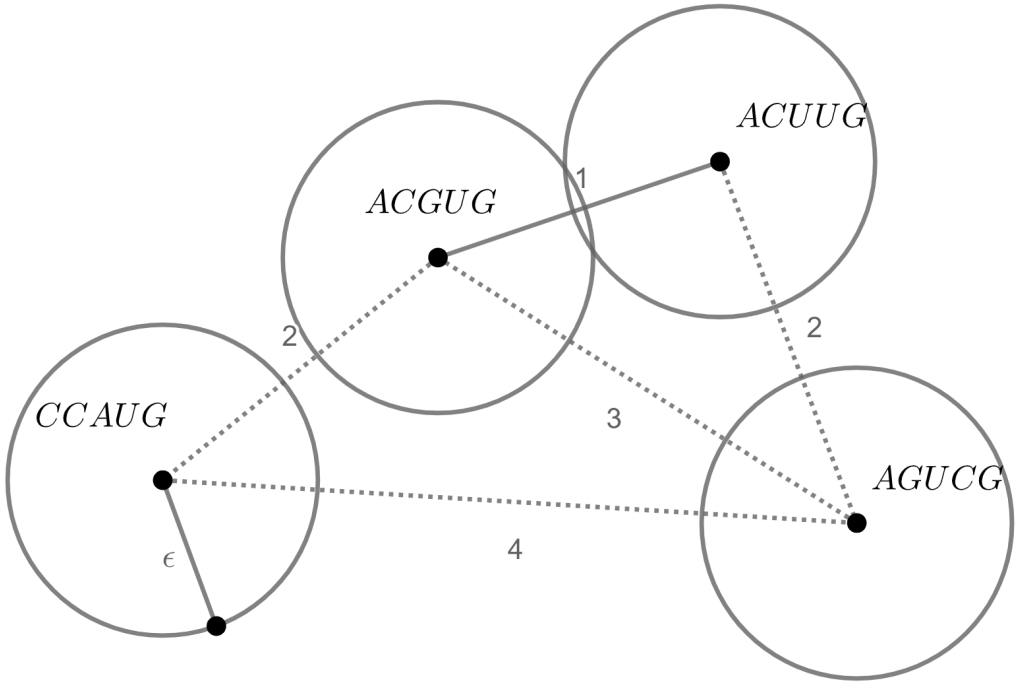


Figure: A 0-simplex, 1-simplex, and 2-simplex.

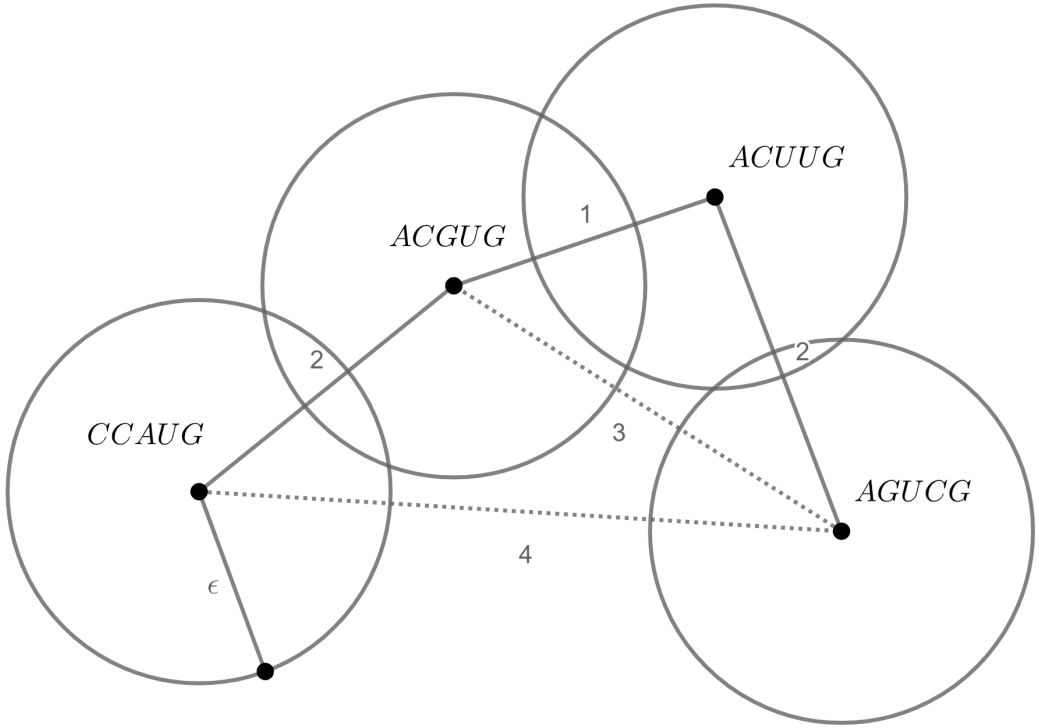
Vietoris-Rips Complex



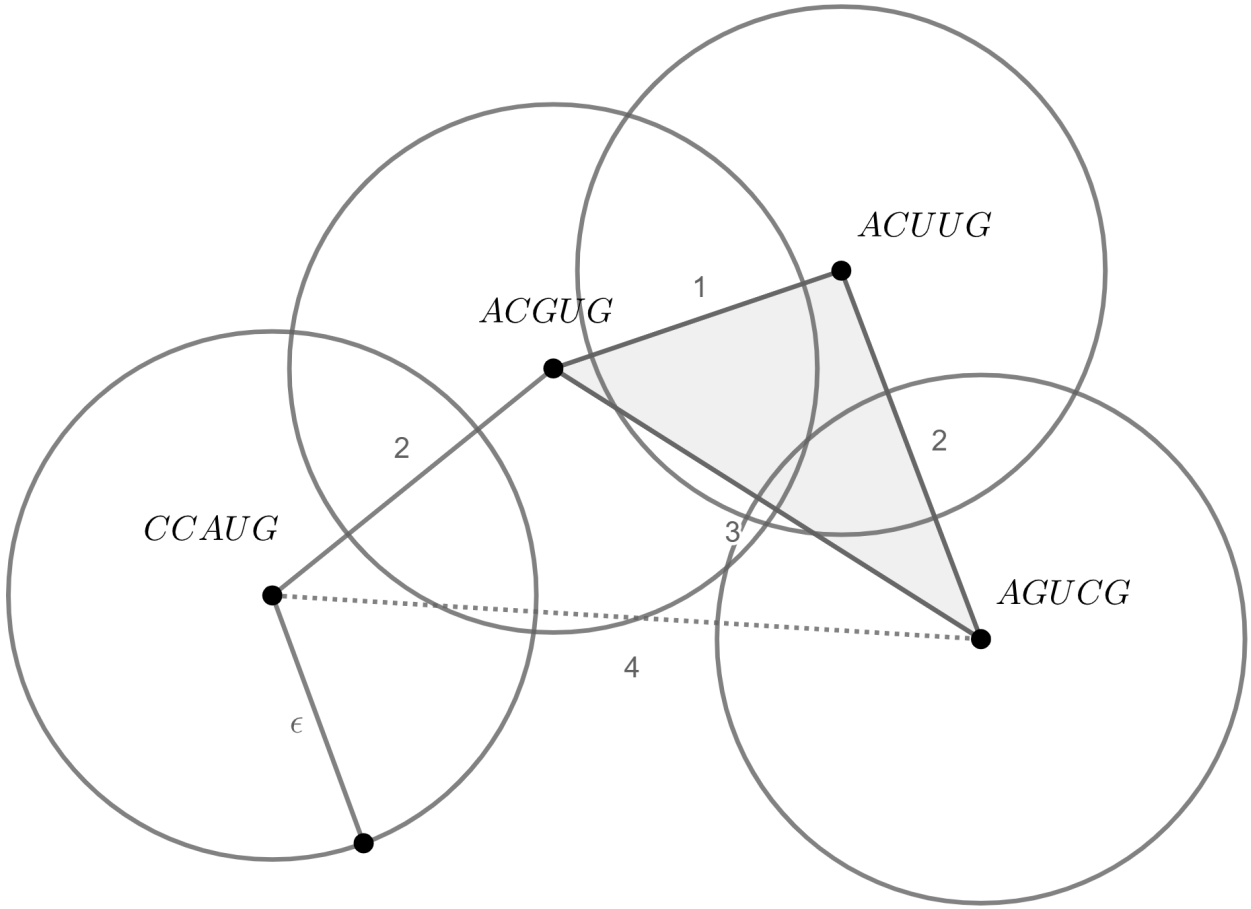
Vietoris-Rips Complex



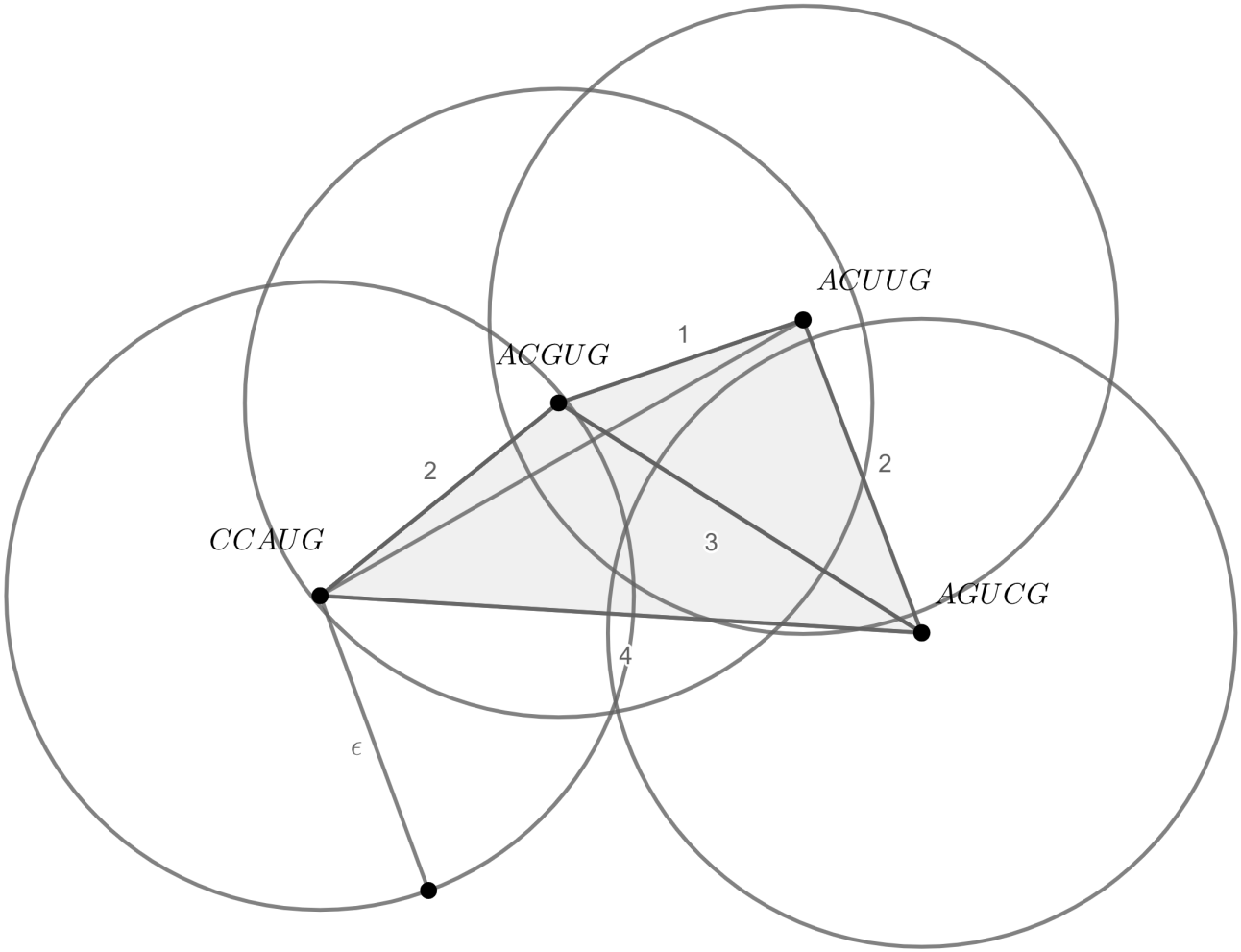
Vietoris-Rips Complex



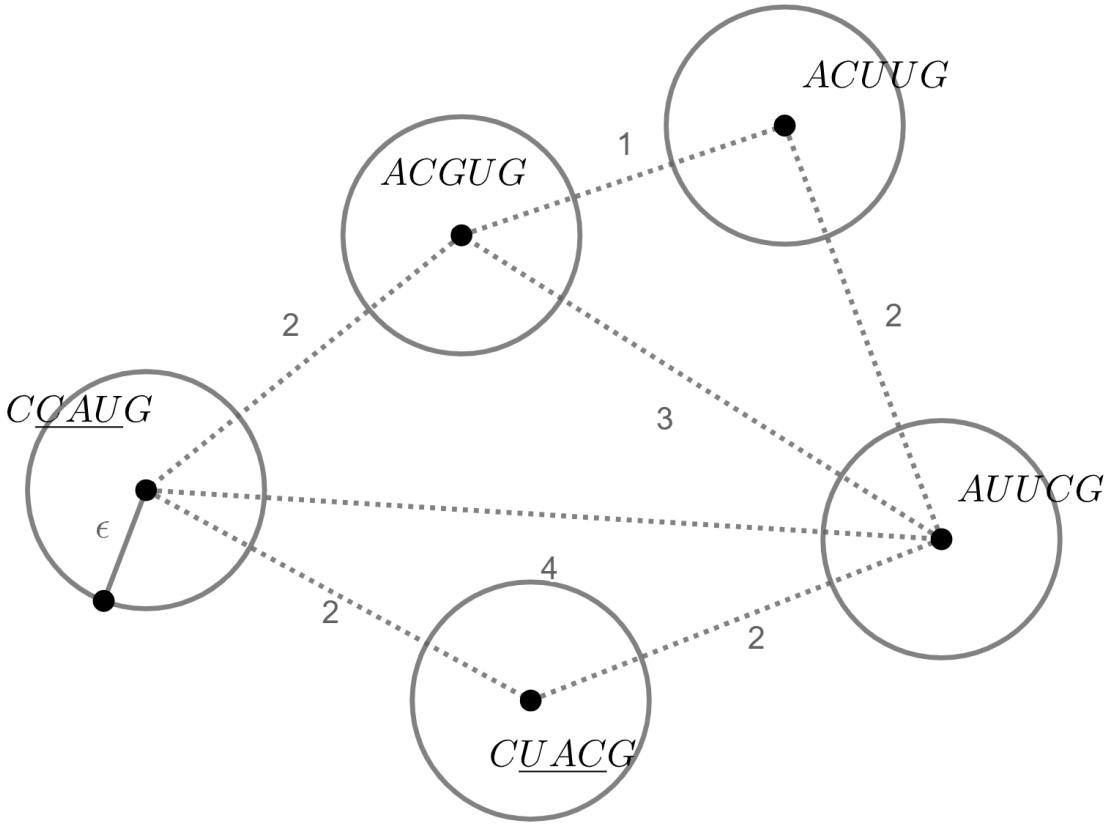
Vietoris-Rips Complex



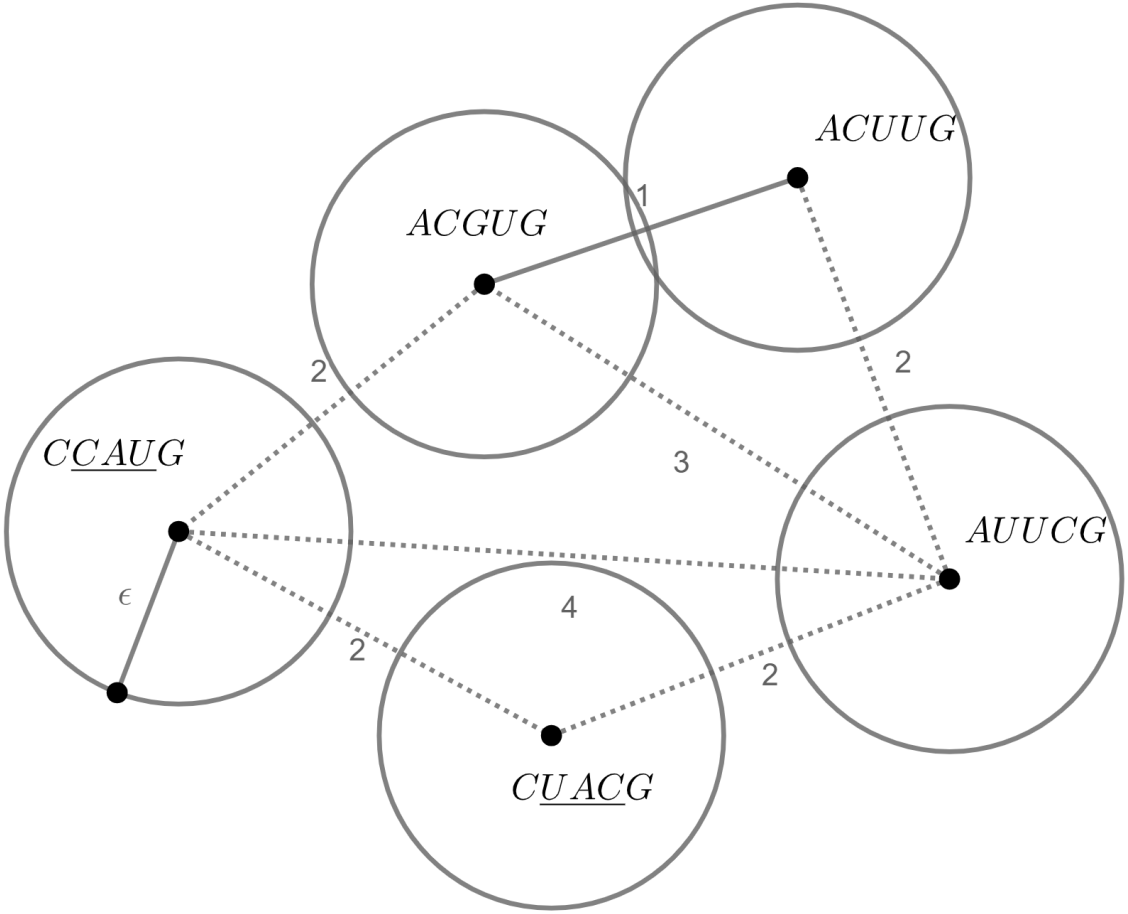
Vietoris-Rips Complex



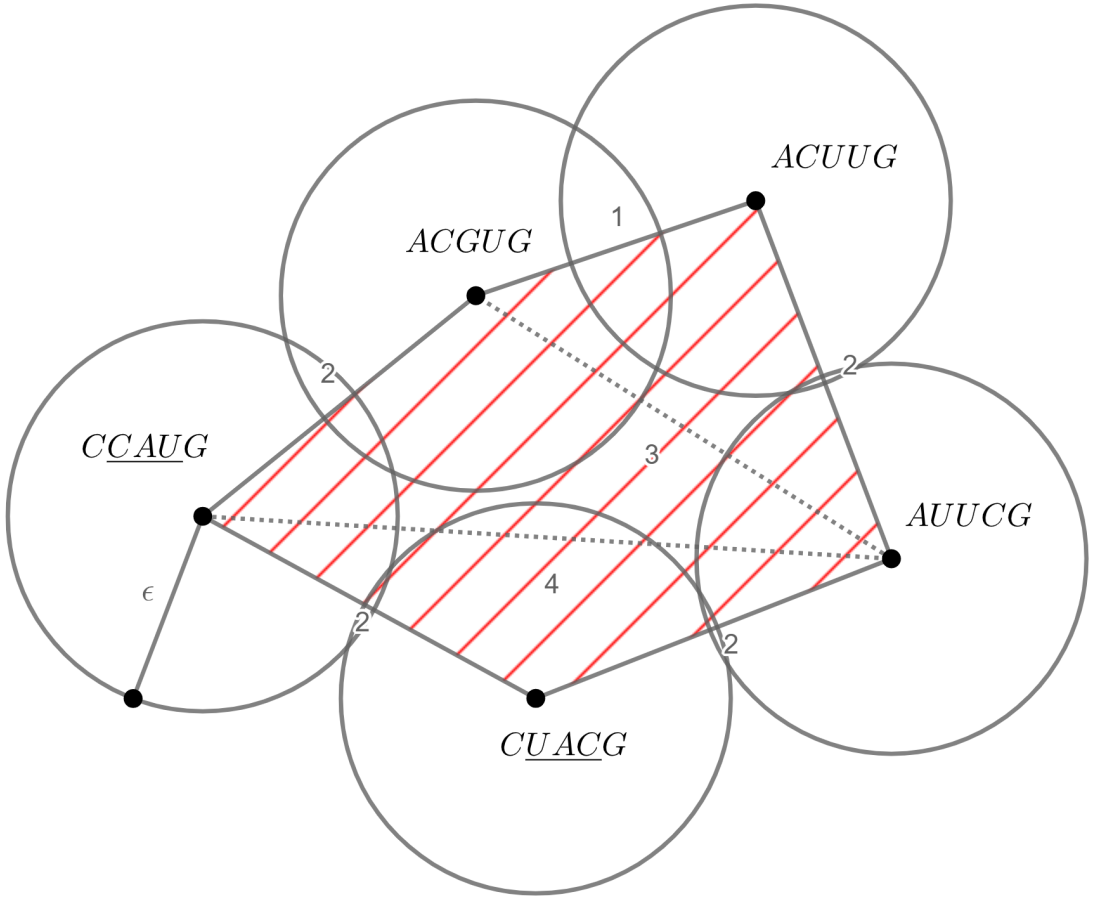
Recombinations Form Irreducible Cycles



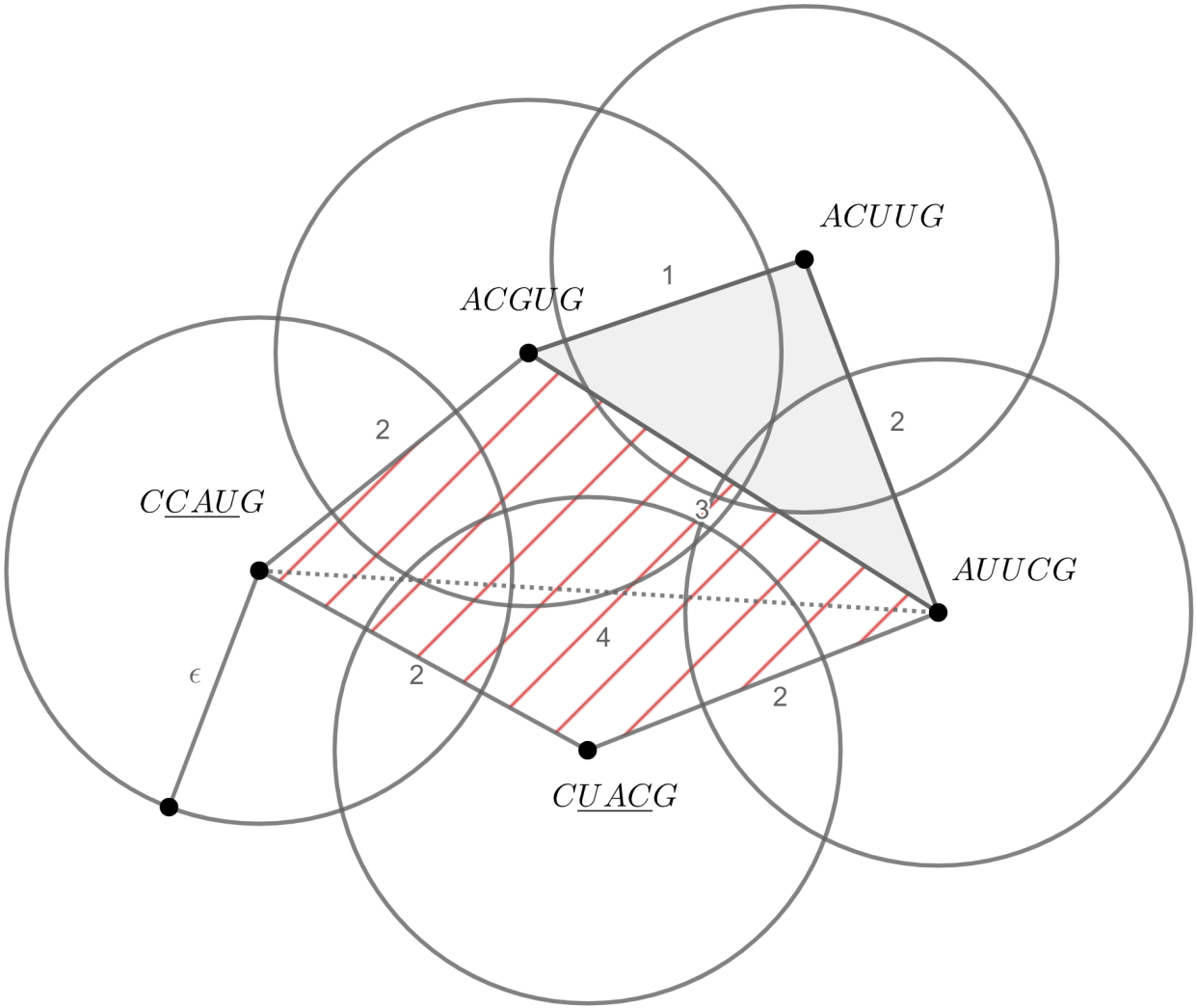
Recombinations Form Irreducible Cycles



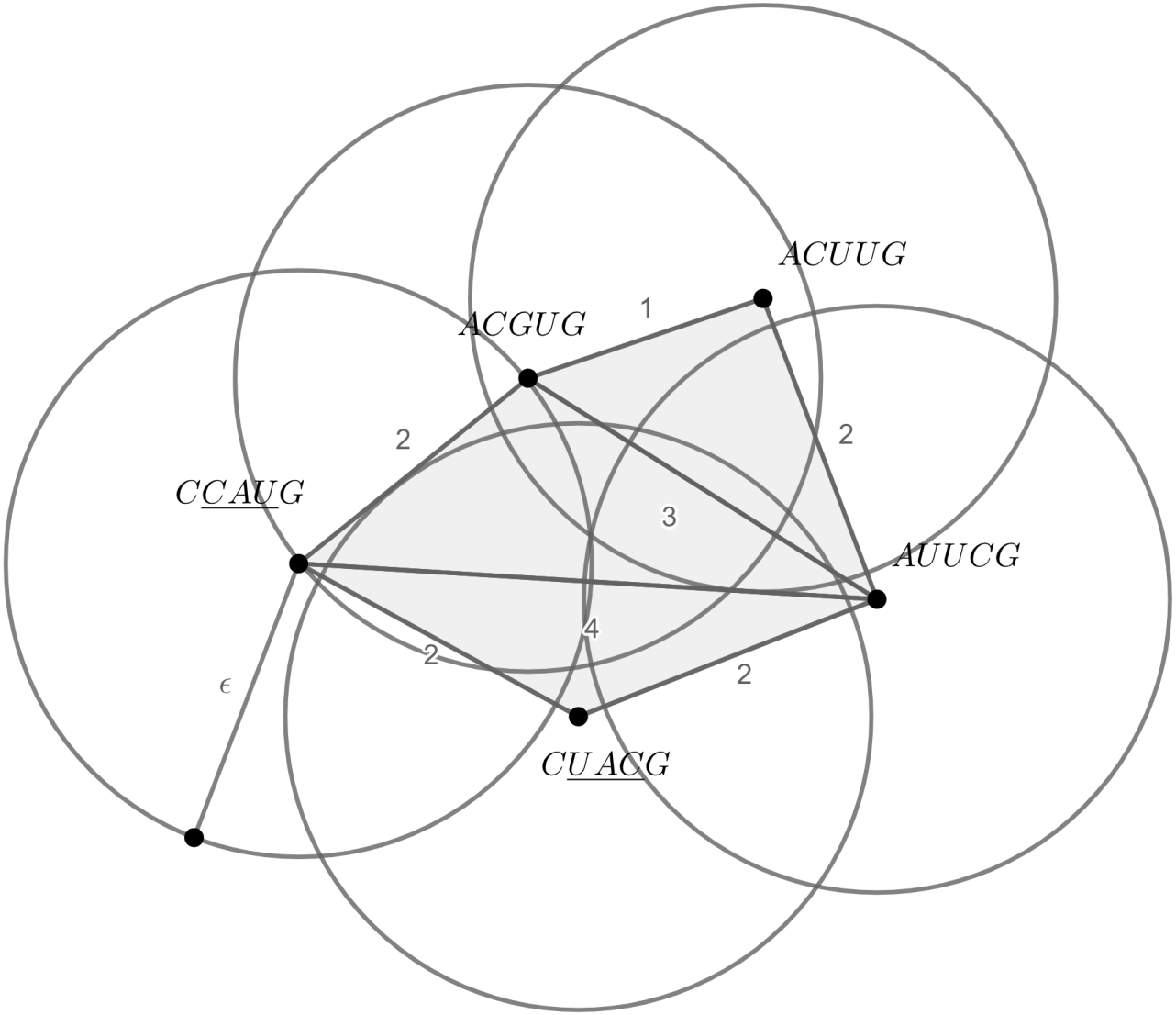
Recombinations Form Irreducible Cycles



Recombinations Form Irreducible Cycles



Recombinations Form Irreducible Cycles



Persistent Homology

As *filtration parameter* ϵ increases, a sequence of simplicial complexes are generated.

Persistent Homology

As *filtration parameter* ϵ increases, a sequence of simplicial complexes are generated.

The *persistence* of a 1-dimensional cycle (hole) is the difference between the maximum and minimum ϵ where it exists in the resulting simplicial complex.

Project Goal

The goal of my project is to use computer simulations to analyze the effectiveness of Topological Data Analysis (TDA) in detecting recombination events.

Project Goal

The goal of my project is to use computer simulations to analyze the effectiveness of Topological Data Analysis (TDA) in detecting recombination events.

Variables of interest:

- Distance metric
- Type of recombination
- Measure of TDA

Simulation Details

100 copies of a single random 1000 nucleotide sequence are generated.

Simulation Details

100 copies of a single random 1000 nucleotide sequence are generated.

Pointwise mutations and recombination events are simulated over 30 generations.

Simulation Details

100 copies of a single random 1000 nucleotide sequence are generated.

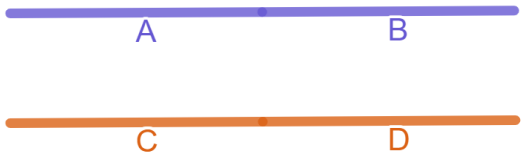
Pointwise mutations and recombination events are simulated over 30 generations.

TDA is run on the resulting sequences with a chosen distance metric and a summary of the result is produced.

Types of Recombination Modeled

Deletion

Insertion



Translocation

Inversion

Distance Metrics on Deletions

Standard Hamming Distance:

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

Distance Metrics on Deletions

Standard Hamming Distance:

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

MEGA-X Hamming Distance:

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

Distance Metrics on Deletions

Standard Hamming Distance:

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

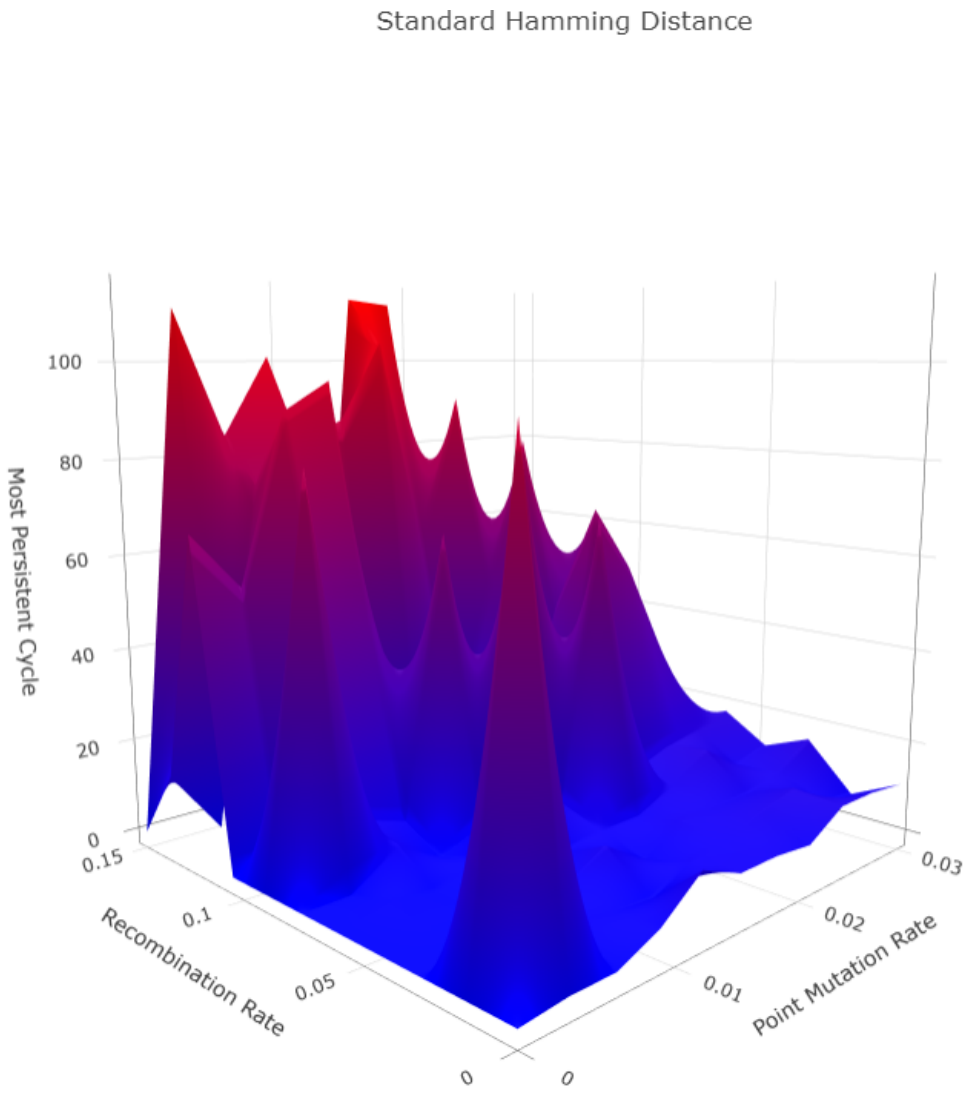
MEGA-X Hamming Distance:

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

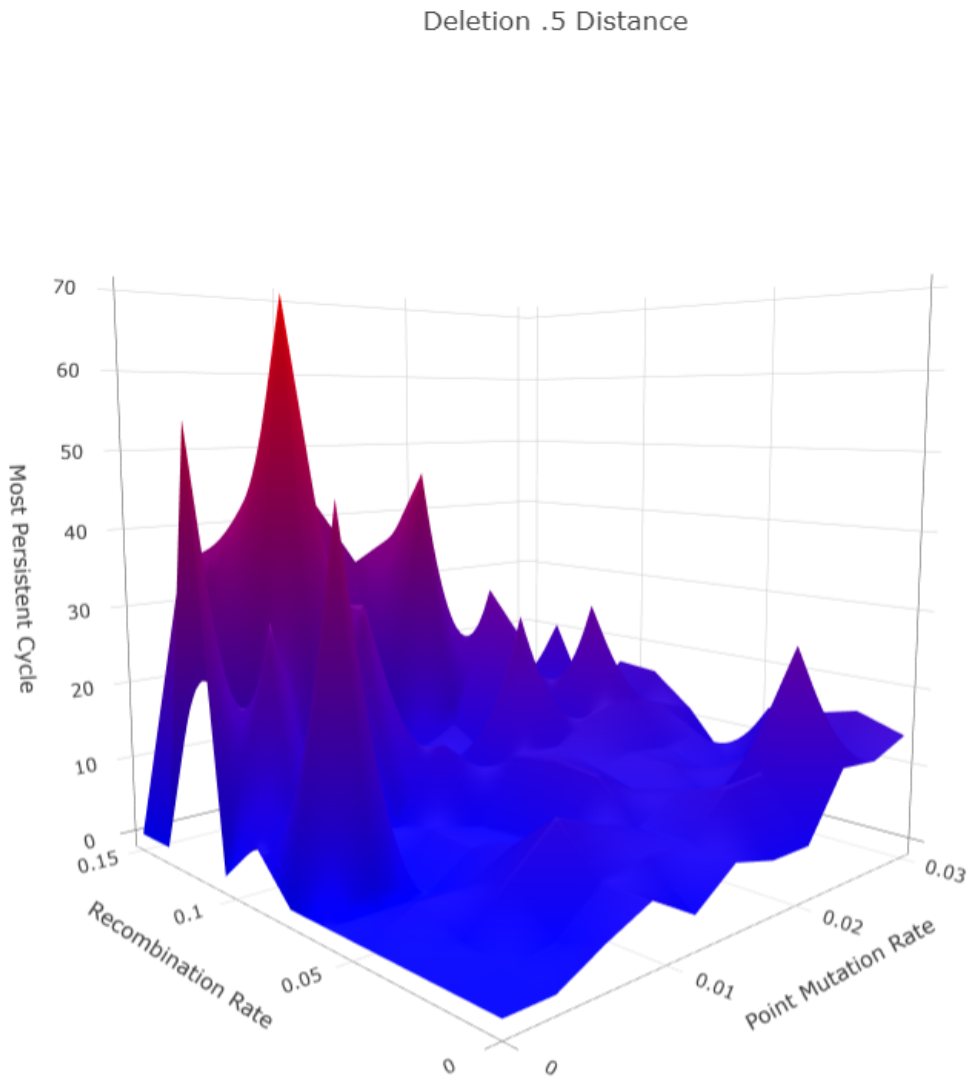
Proposed Distance:

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

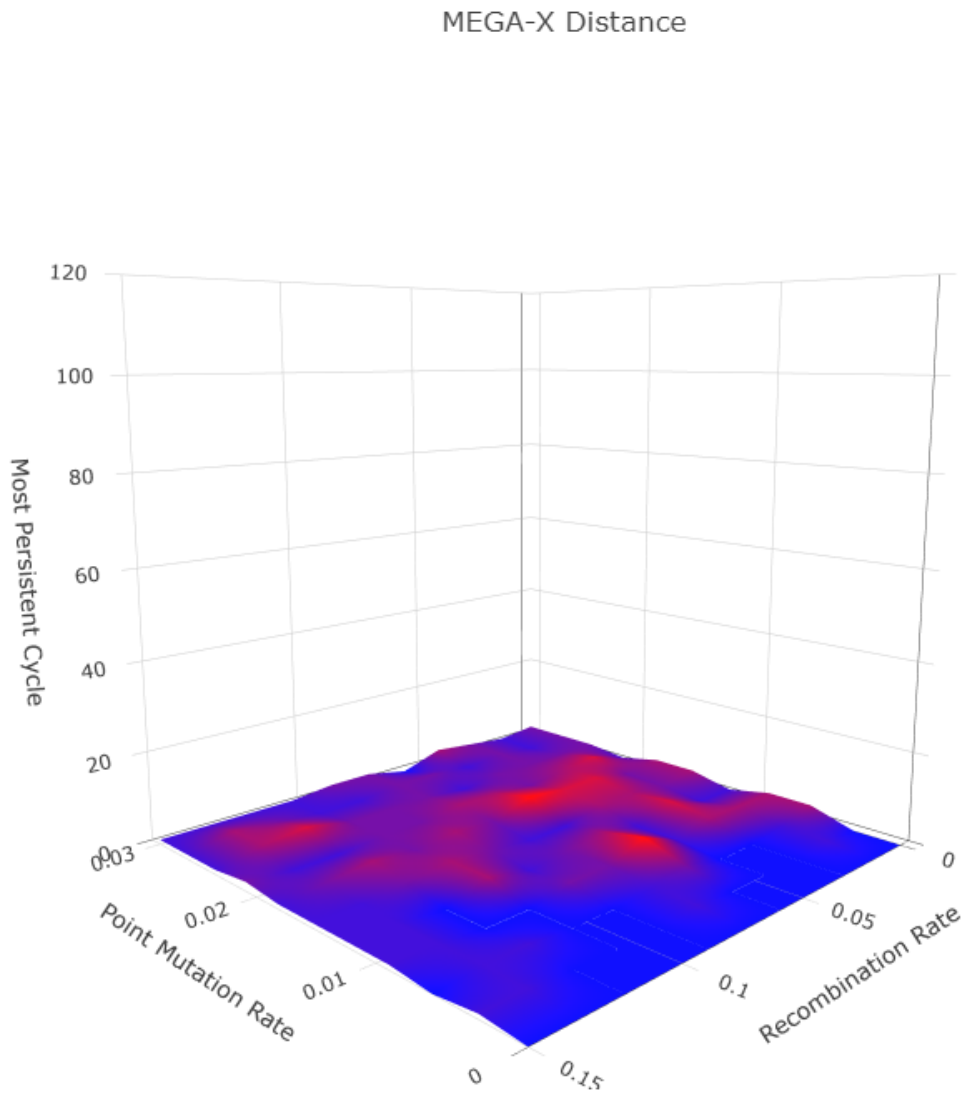
Standard Hamming Distance on Deletions



Proposed Distance on Deletions



MEGA-X on Deletions



Distance Metric: Conclusions

Since the distance metrics given only differ on deletions, they should perform the same on translocation and inversion.

Distance Metric: Conclusions

Since the distance metrics given only differ on deletions, they should perform the same on translocation and inversion.

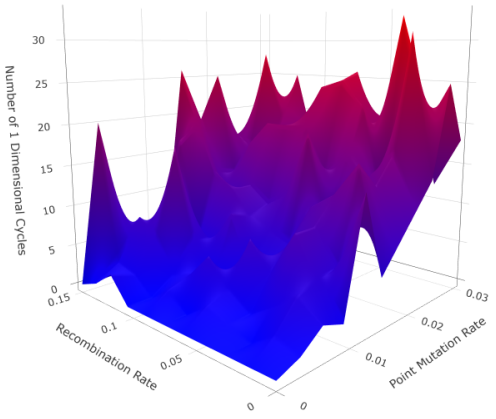
The standard Hamming distance detected deletions more effectively than the proposed metric, and the MEGA-X metric performed by far the worst.

Measures of TDA

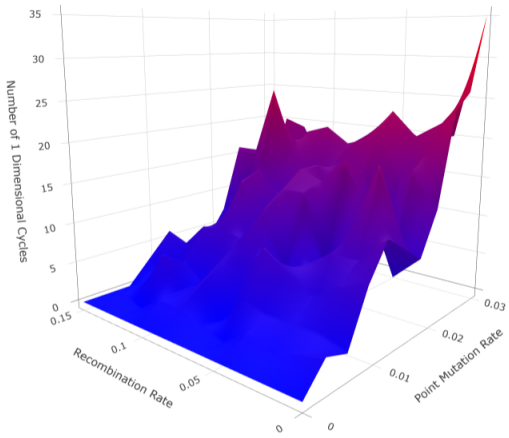
- Number of 1-dimensional cycles.
- Maximum persistence.
- Sum of cycles' persistence.

Number of Cycles

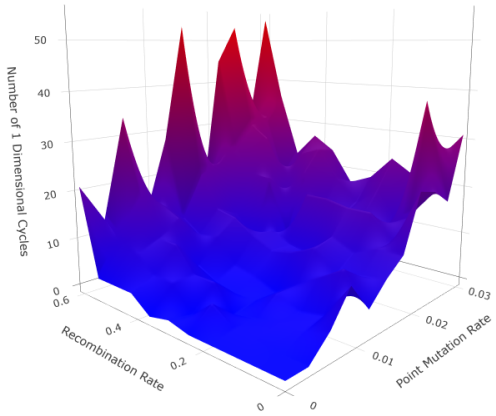
Deletions



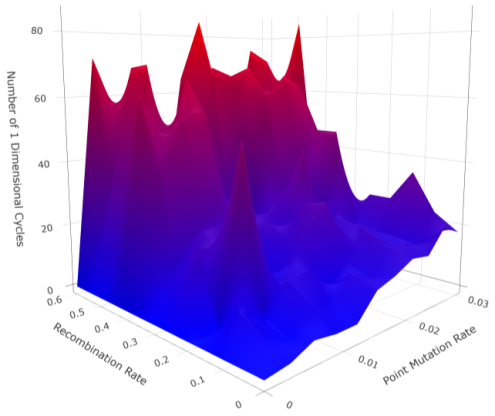
Insertions



Inversions

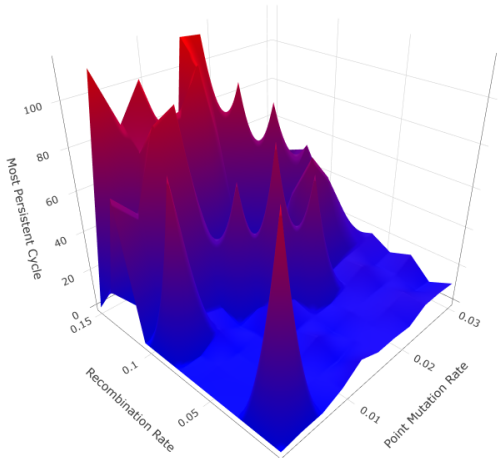


Translocations

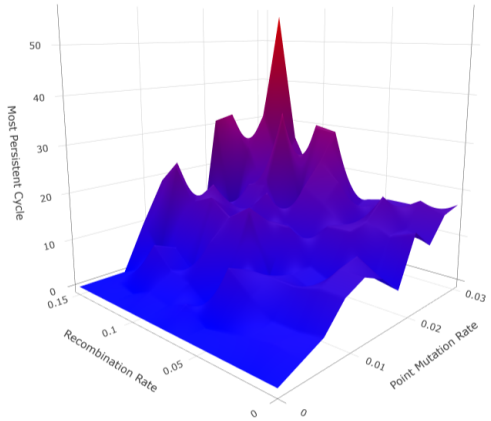


Most Persistent Cycle

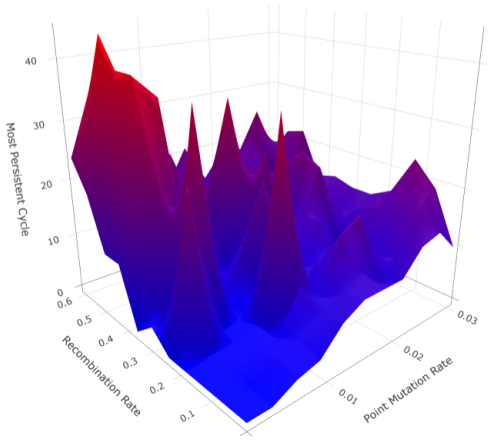
Deletions



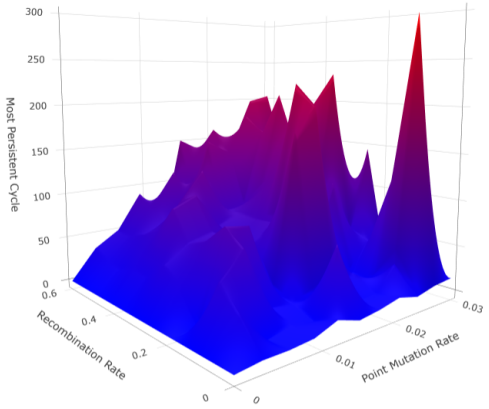
Insertions



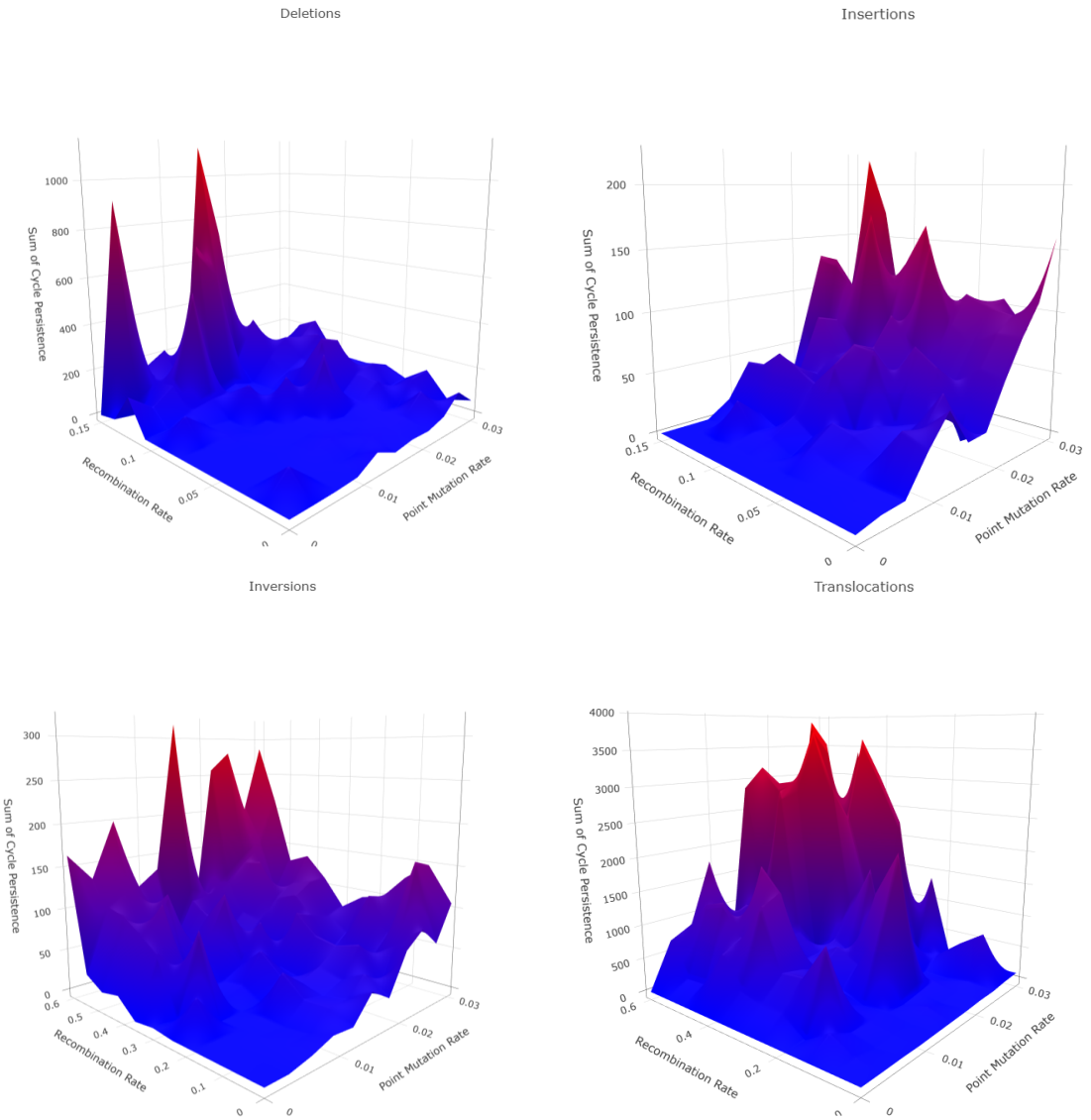
Inversions



Translocations



Sum of Persistence Levels



Measure of TDA: Conclusions

Maximum persistence was the most effective for each type of recombination except translocation, but had high variation (many spikes).

Measure of TDA: Conclusions

Maximum persistence was the most effective for each type of recombination except translocation, but had high variation (many spikes).

Number of 1-dimensional cycles was ineffective, except for translocation, where it was effective.

Measure of TDA: Conclusions

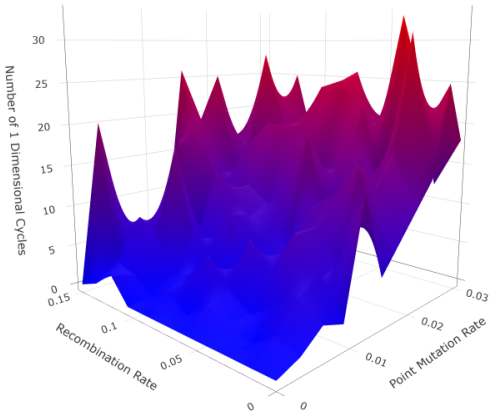
Maximum persistence was the most effective for each type of recombination except translocation, but had high variation (many spikes).

Number of 1-dimensional cycles was ineffective, except for translocation, where it was effective.

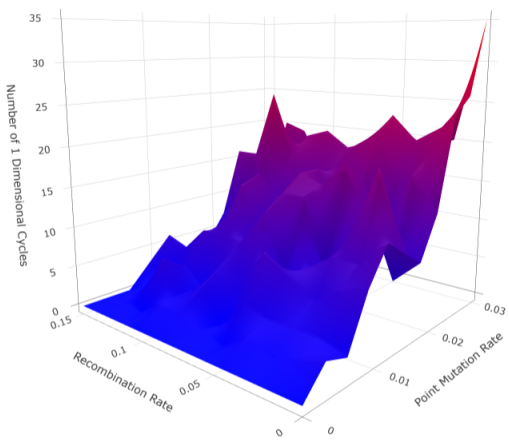
The sum of the cycle's persistence tended to lie between the maximum persistence and number of 1-dimensional cycles. This led it to be marginally effective for each case.

Types of Recombination: Number of Cycles

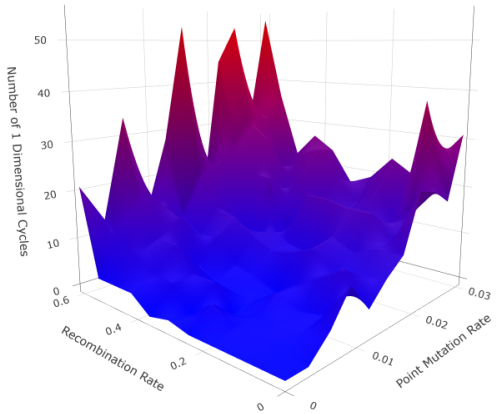
Deletions



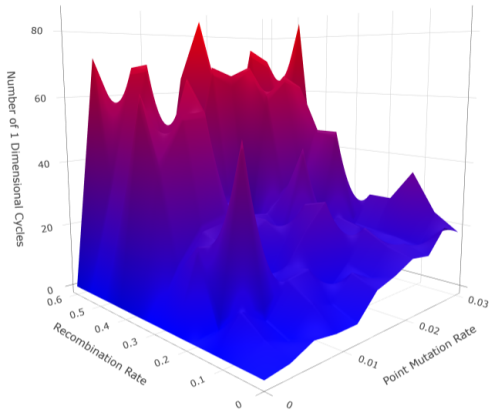
Insertions



Inversions

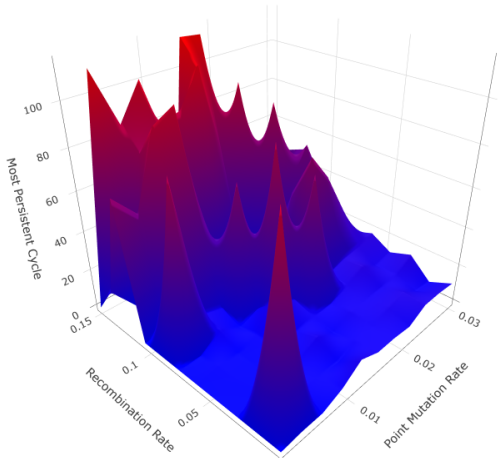


Translocations

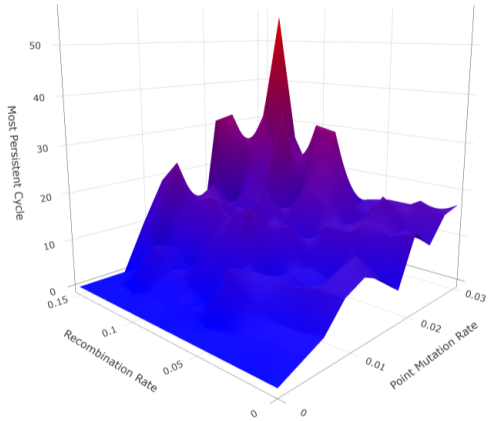


Types of Recombination: Maximum Persistence

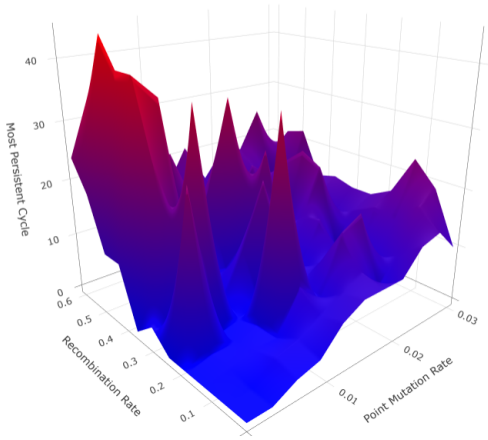
Deletions



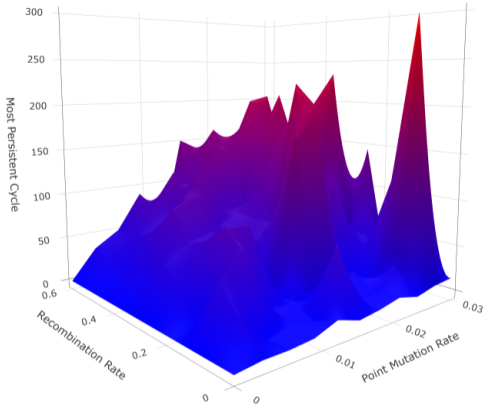
Insertions



Inversions



Translocations



Type of Recombination: Conclusions

Insertions were the most difficult for TDA to detect, while inversions and deletions were well detected.

Type of Recombination: Conclusions

Insertions were the most difficult for TDA to detect, while inversions and deletions were well detected.

Translocations were well detected by the number of cycles, but was not detected well by maximum persistence.

Further Questions:

- Alpha complex in discrete space.

Further Questions:

- Alpha complex in discrete space.
- Prove the "spikiness" of the maximum persistence.

Thank You!!!

- Dr. Javier Arsuaga and Dr. Máriel Vazquez
- Kristina Moen, Sofia Jakovcevic, Michael Keith
- Everyone in the Biophysics and BioMath groups
- All the REU staff and students that made this summer so great!