

Topological Data Analysis with Mapper: an Implementation in Cytoscape and an Application to Aptamers

George Clare Kennedy

August 29, 2024

University of Iowa

Outline

Why TDA?

Mapper and Its Flavors

Cytoscape to the Rescue

Aptamers

Mapping out Future Directions

Section Map (!)

Why TDA?

Mapper and Its Flavors

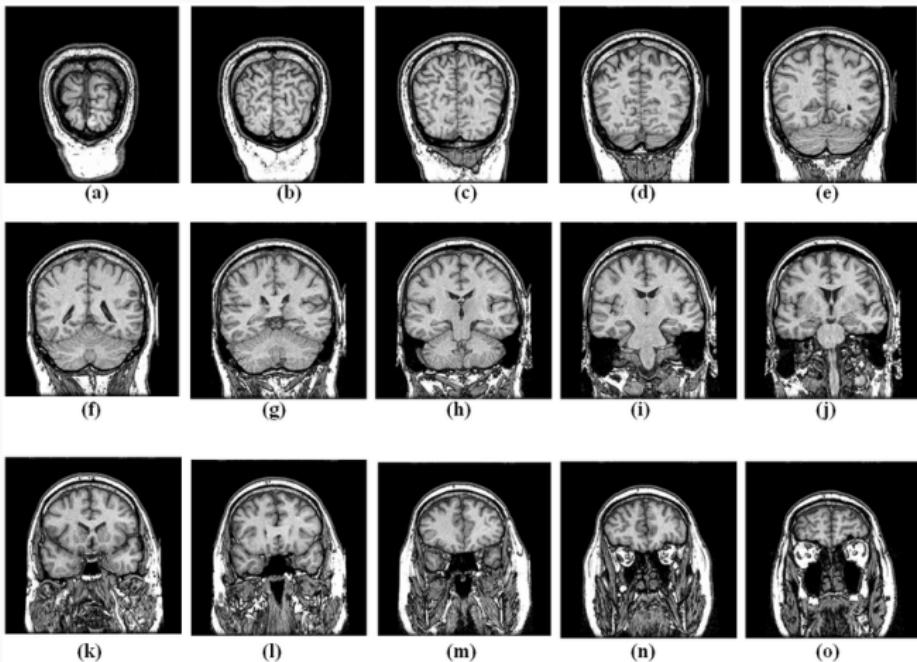
Cytoscape to the Rescue

Aptamers

Mapping out Future Directions

Data is Big

- Modern techniques allow for rich data collection and storage
- Size of datasets can be enormous in both observations (rows) and variables (columns)



Data is Big

- Modern techniques allow for rich data collection and storage
- Size of datasets can be enormous in both observations (rows) and variables (columns)



Geometry is Hard

- High-dimensional space is extremely unintuitive
- If $V_n(r)$ is the volume of the n -dimensional ball with radius r , then for any $\varepsilon > 0$,

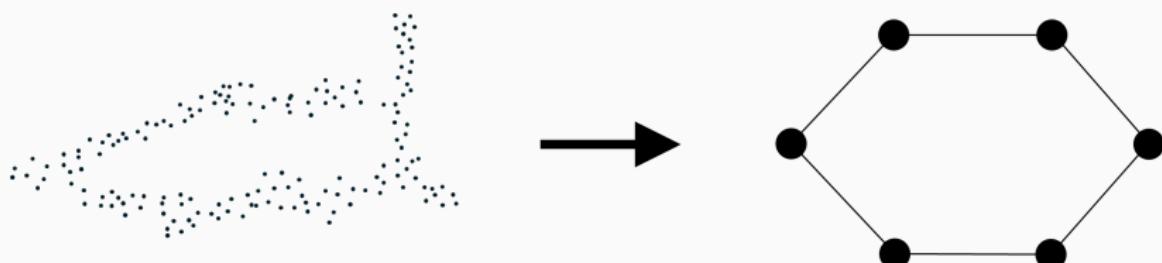
$$\lim_{n \rightarrow \infty} \frac{V_n(1 - \varepsilon)}{V_n(1)} = 0$$

i.e., the volume of balls lives almost entirely at the boundary

- Trying to analyze many characteristics creates combinatorial problems ($n!$ is big!)

Toning It Down

- Broad idea: high dimensions \implies low dimensions
- More specific idea: build a simplicial complex
- Simpler idea: build a 1-dimensional simplicial complex (that is, a graph)
- Enter: the Mapper algorithm (Singh et al, 2007)



Section Map (!)

Why TDA?

Mapper and Its Flavors

Cytoscape to the Rescue

Aptamers

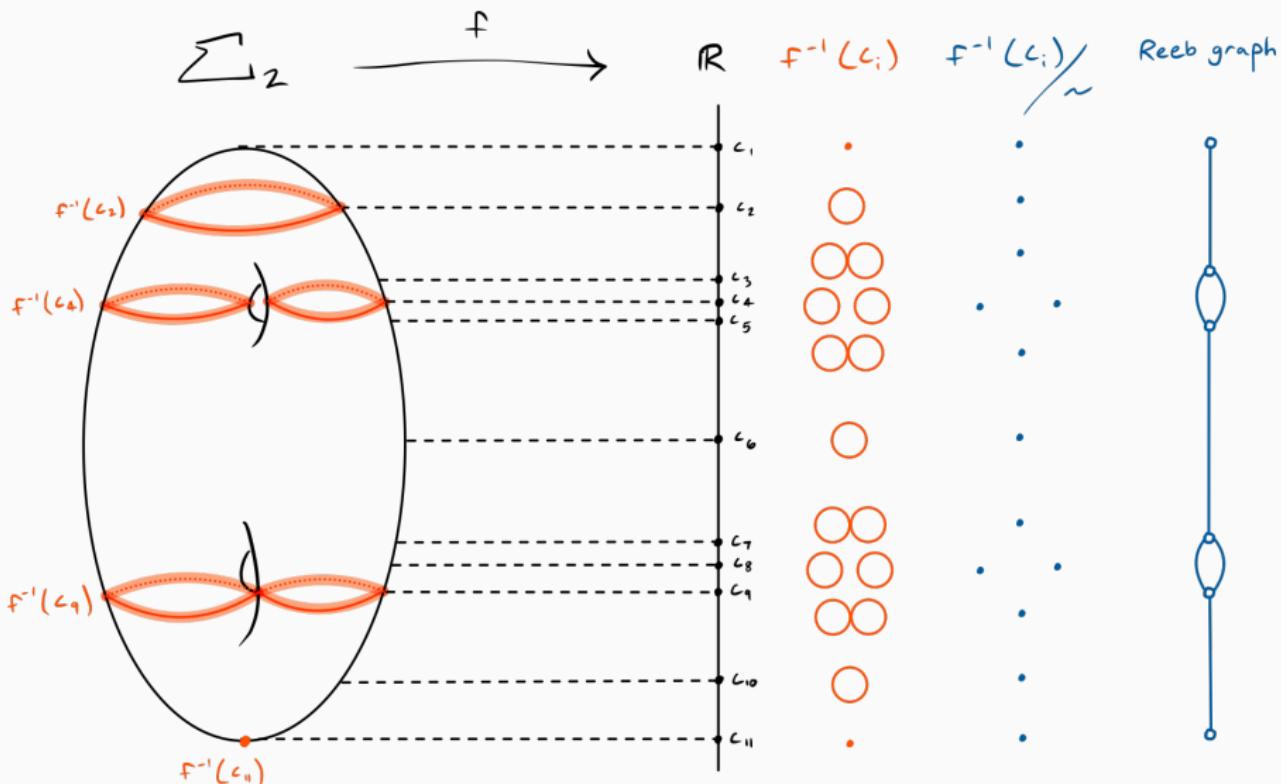
Mapping out Future Directions

Motivation: Reeb Graph

- Idea: construct graph reflecting level sets of a “filter” function
- Formally, given a topological space X and a continuous function $f : X \rightarrow \mathbb{R}$, define an equivalence relation \sim on X where $x \sim y$ if x and y live in the same connected component of a level set $f^{-1}(c)$ for some $c \in \mathbb{R}$.
- The **Reeb graph**¹ is X / \sim , taken with the quotient topology.

¹Despite names this is not always a graph

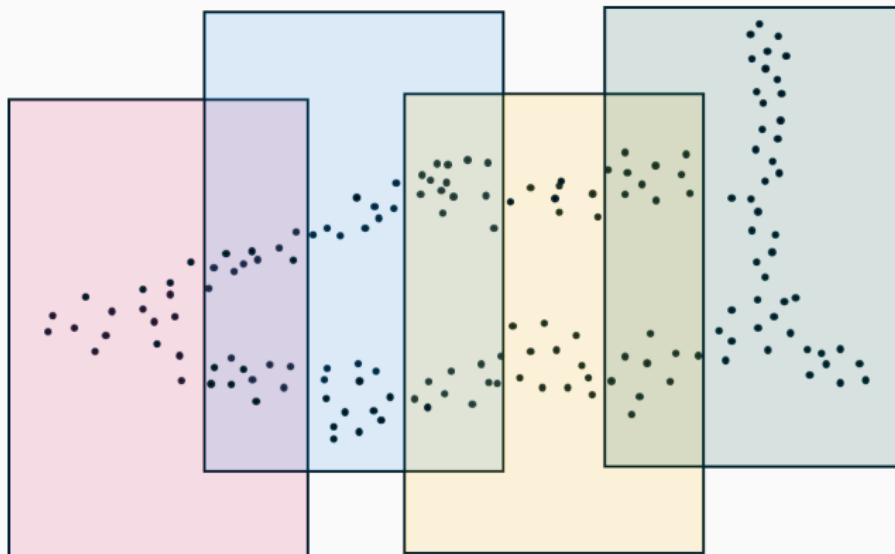
Motivation: Reeb Graph



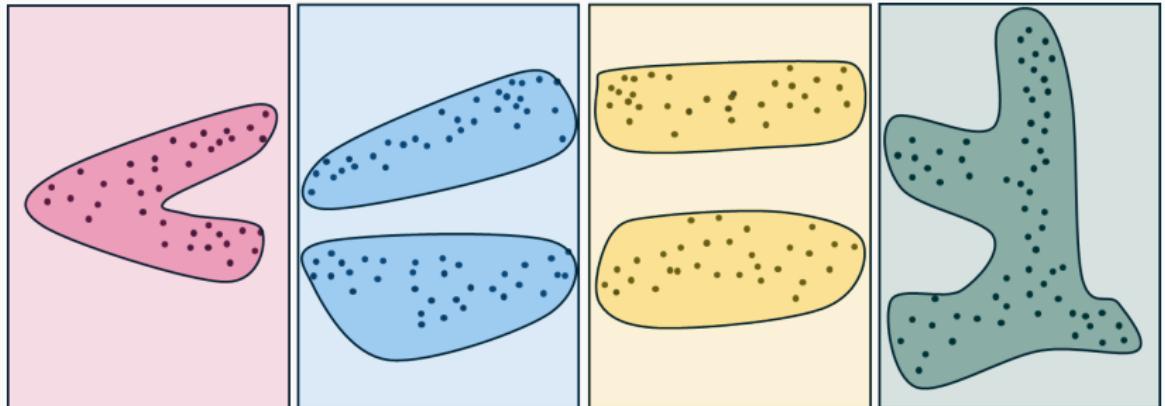
Mapper: Original Flavor

- How can we apply this to the discrete setting?
- Topological space $X \implies$ point cloud P (a discrete set of points in a space)
- Filter function: $f : P \implies \mathbb{R}$
- Level sets of points \implies level sets of overlapping intervals
- Connected components \implies clusters
- Quotient space \implies nerve complex

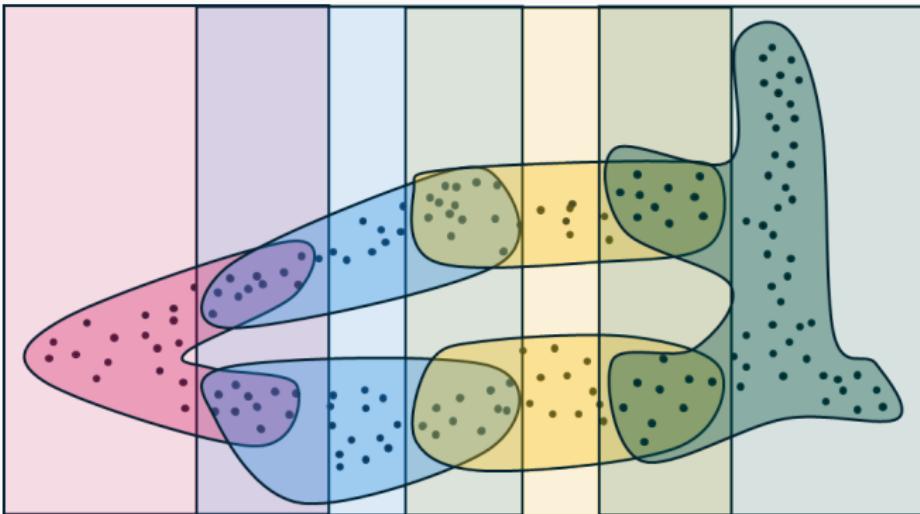
Filtering



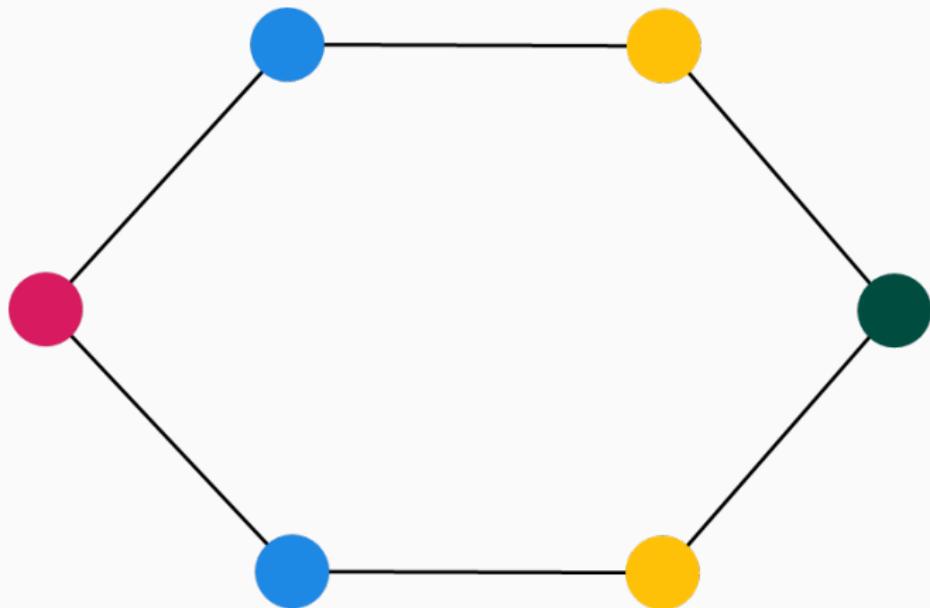
Clustering



Nerve Construction

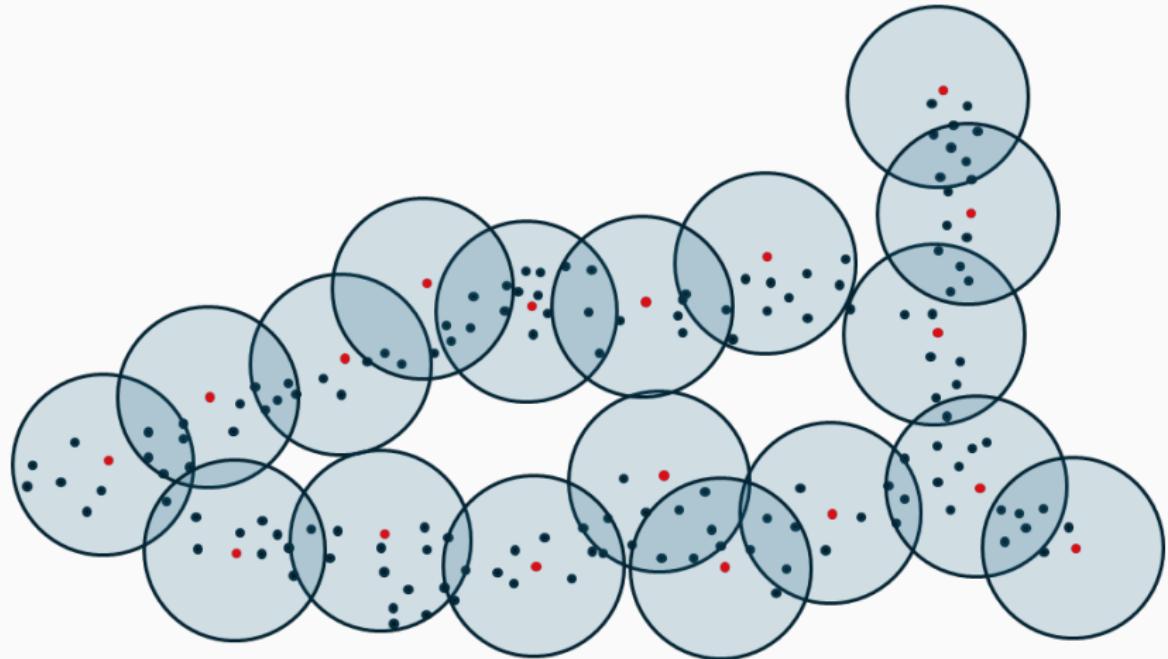


Output

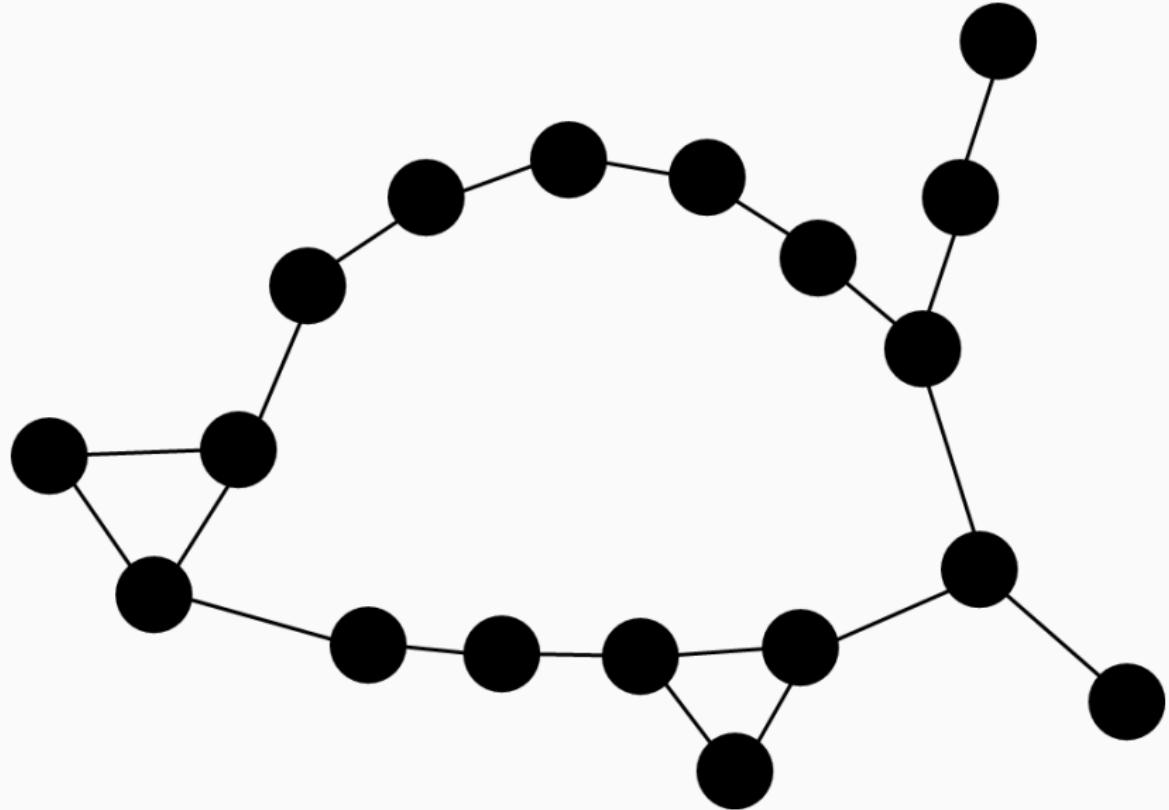


- Original flavor has a lot of choices to make
- Idea: come up with a one-parameter Mapper
- Ballmapper: in place of a conventional filter, cover the dataset with overlapping ε -balls
- Specifically, we want a cover $C = \bigcup_i B(x_i, \varepsilon)$ such that:
 - Every datapoint x is contained in some $B(x_i, \varepsilon)$ for some x_i
 - If x_j is a ball center, then the only ball containing it is $B(x_j, \varepsilon)$
- Nerve construction is the same

Ballmapper: Cover

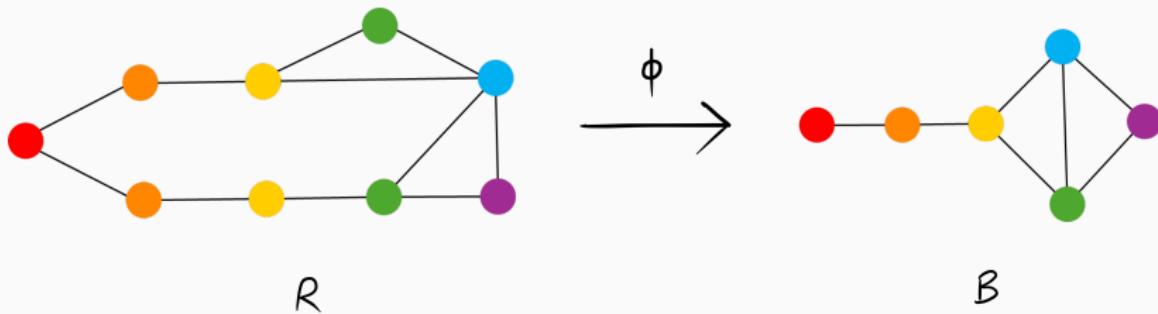


Ballmapper: Nerve

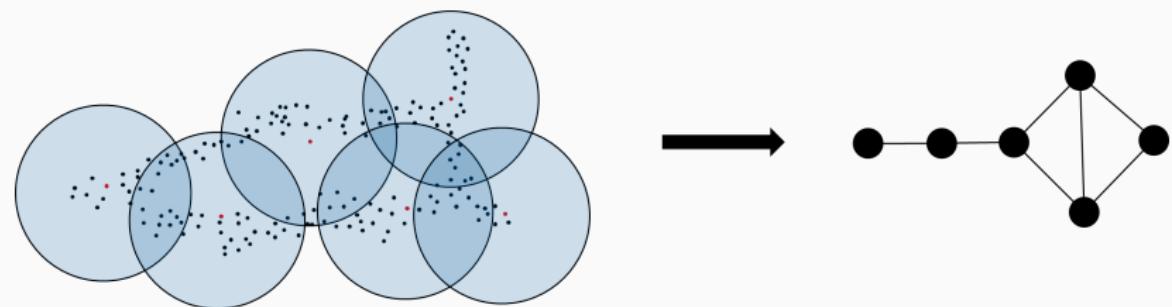


Refined Ballmapper

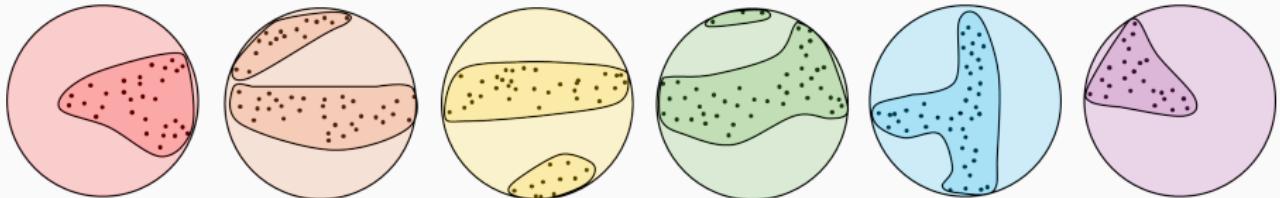
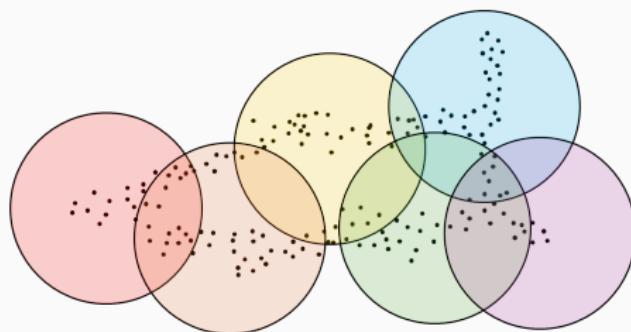
- Idea: combine Ball filtering and Original clustering
- Bin by balling, then cluster within balls as in Original
- Allows for comparison of two different metrics at the same time on the same data
- Output graph R is a refinement of Ballmapper output B ; there is a natural graph homomorphism $\phi : R \rightarrow B$ which simply maps vertices into their balls



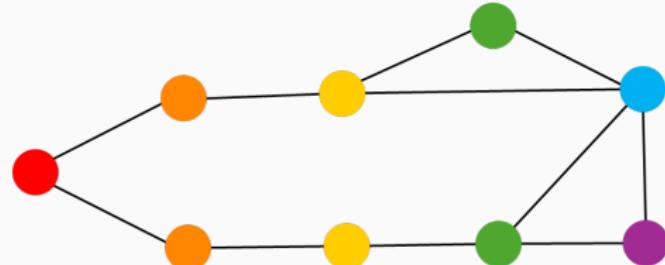
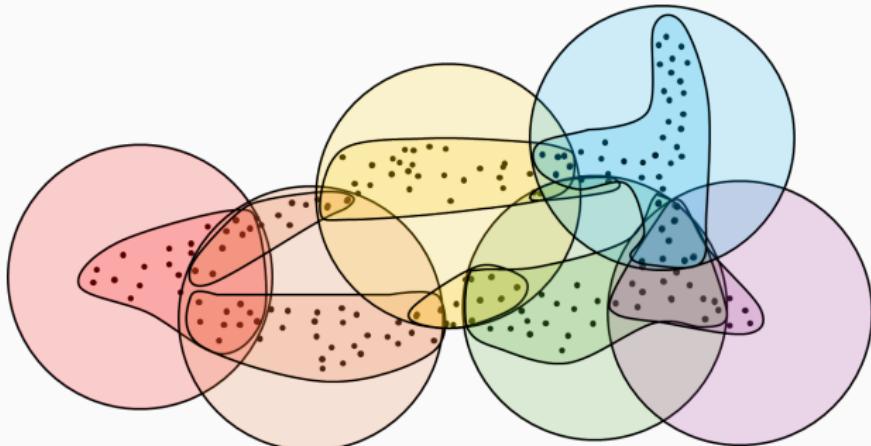
Standard Ballmapper



Balls as Bins

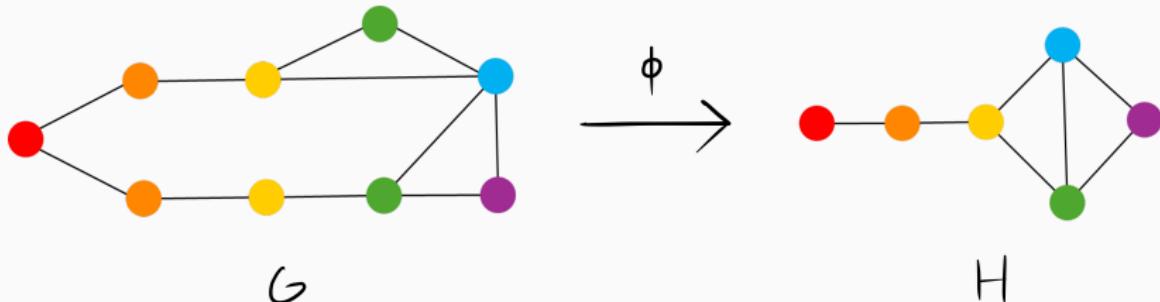


Refined Ballmapper Graph



Graph Theoretic Relation

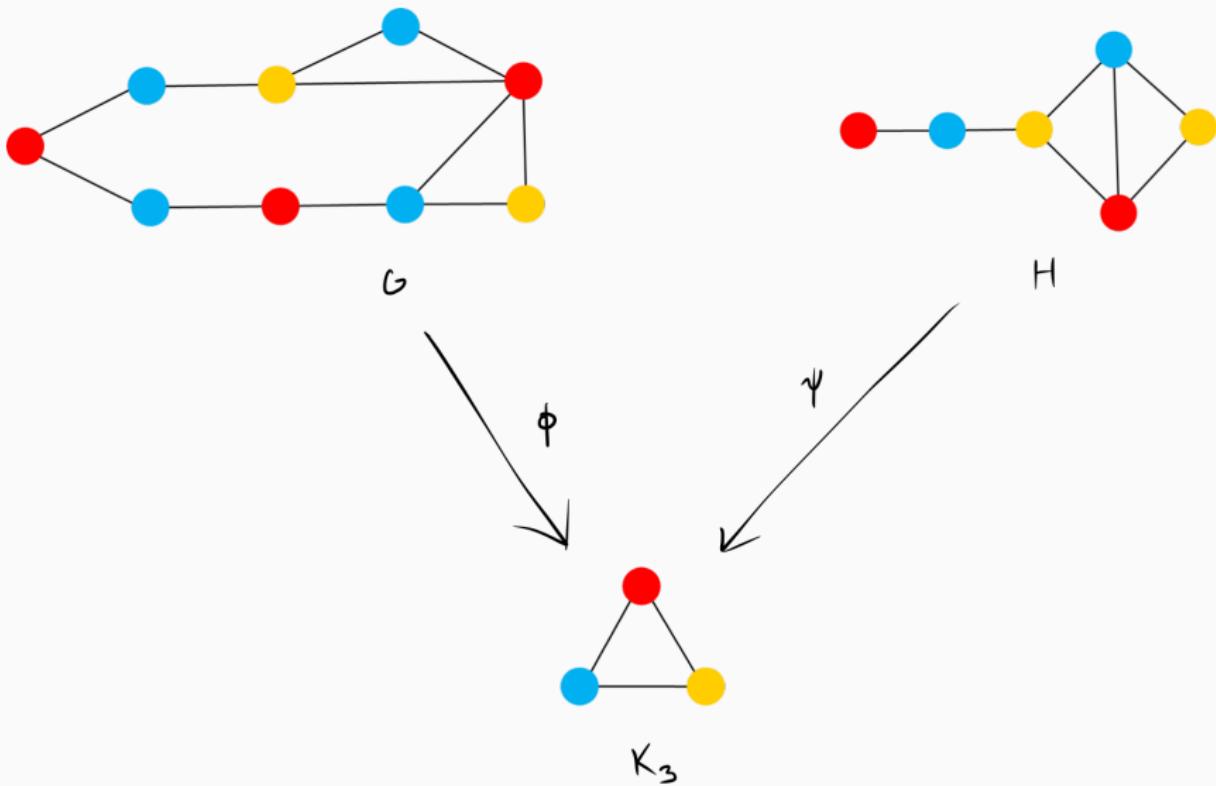
- A function ϕ between the vertices of two graphs G and H is called a **graph homomorphism** if $uv \in E(G)$ implies $\phi(u)\phi(v) \in E(H)$.
- G and H are called **homomorphically equivalent** (hom-equivalent) if there exist graph homomorphisms $f : G \rightarrow H$ and $g : H \rightarrow G$.



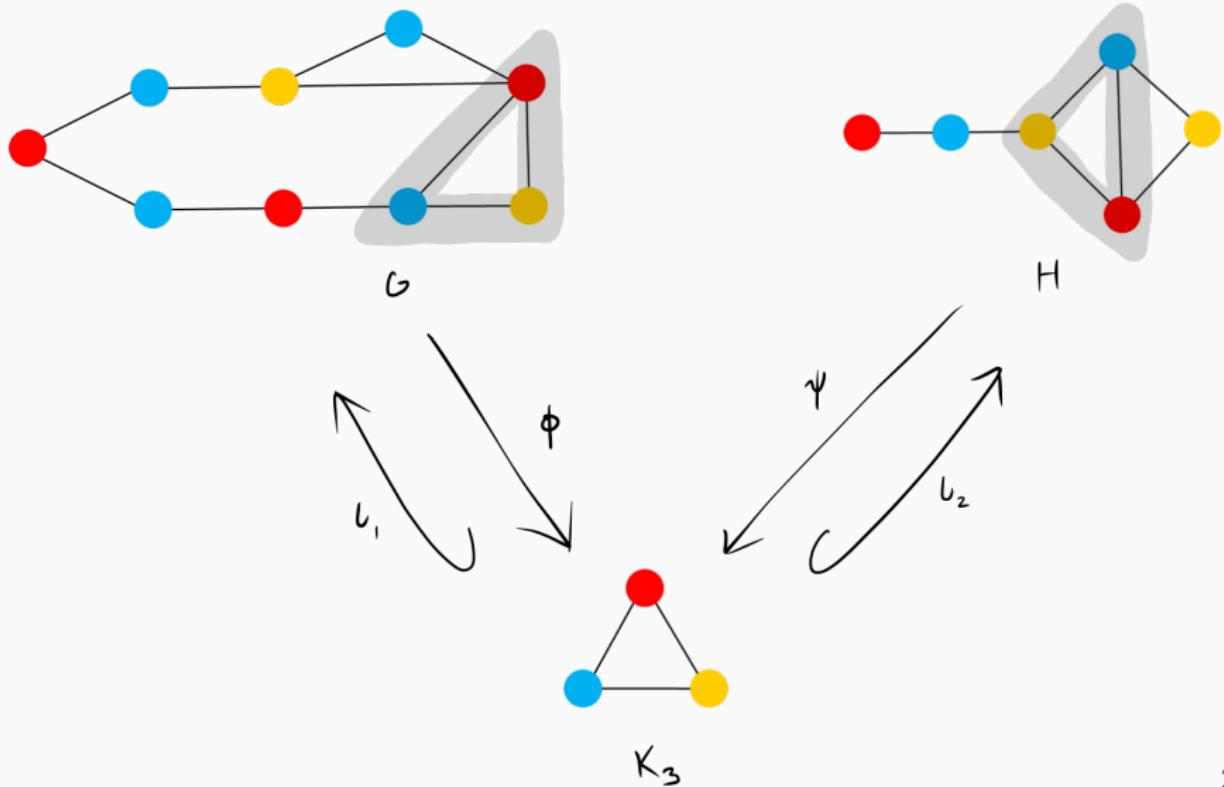
Why Might We Care? Possibility: Cores

- A **core** C of a graph G is a graph such that G and C are hom-equivalent, and C is the smallest such graph.
 - Complete graphs, odd cycles, etc
- Every finite graph has a core, and it is unique (up to isomorphism).
- Graphs with the same cores are necessarily hom-equivalent, and vice versa.
- Core-finding complexity: NP-complete :(
- Applications to relational algebra

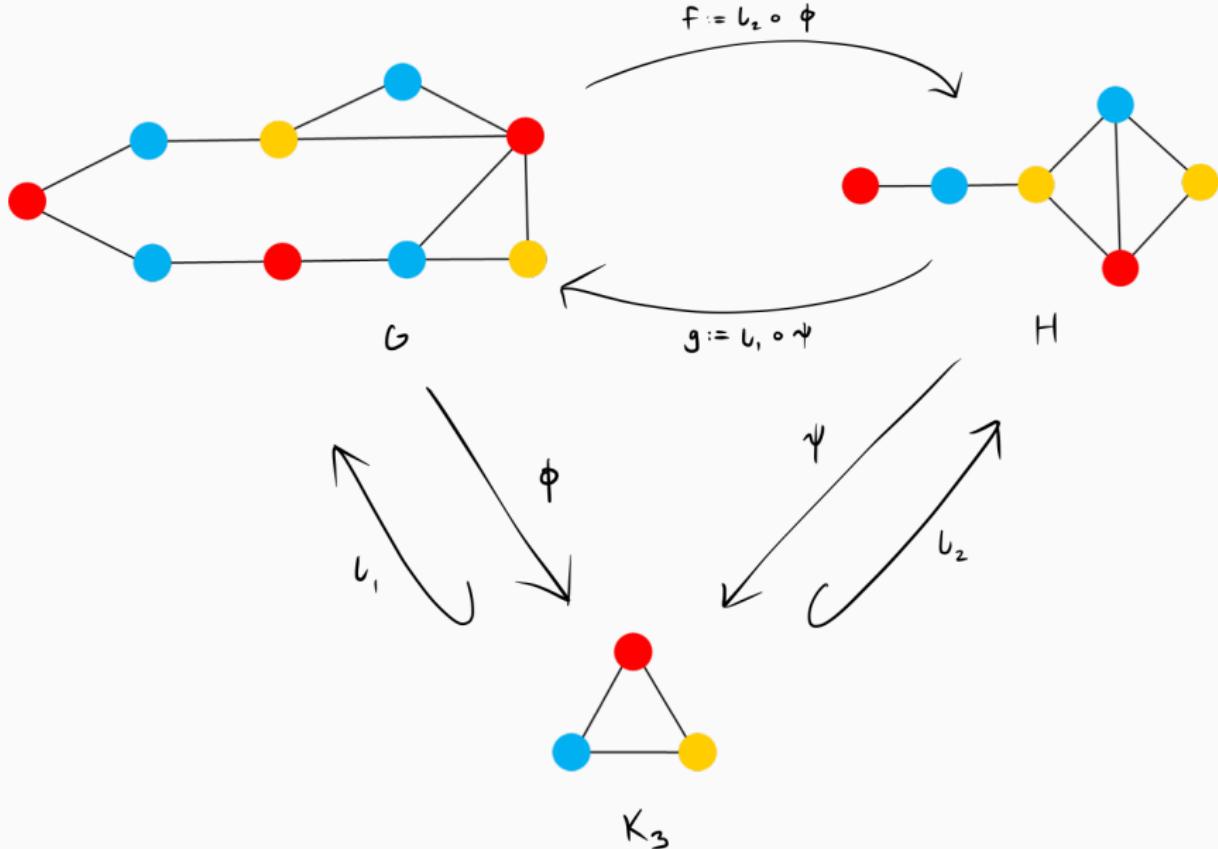
Example Equivalence via Core



Example Equivalence via Core



Example Equivalence via Core



Problems

- We lost so much stuff! That clustering took work...
- Graphs are abstract combinatorial structures; they convey no geometrical information
- Potentially interesting features can “bypass” the filter
- Output heavily depends on your choices of parameters and clustering method

Section Map (!)

Why TDA?

Mapper and Its Flavors

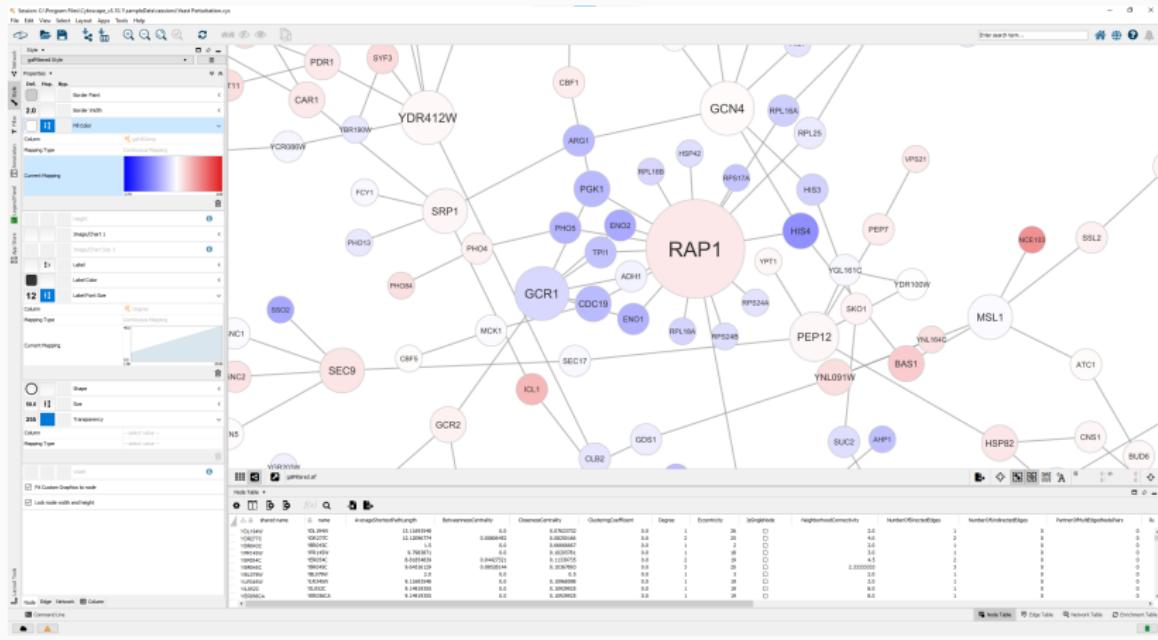
Cytoscape to the Rescue

Aptamers

Mapping out Future Directions

What is Cytoscape?

- Powerful network analysis software (written in Java)
 - Used primarily by bioinformaticists but is a general use program



Features

- Networks are stored as separate tables of nodes and edges
- Tables can be augmented with any number or type of columns
- Any column can then be associated with a visual characteristic of the network

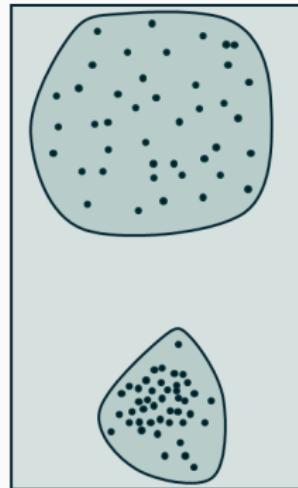
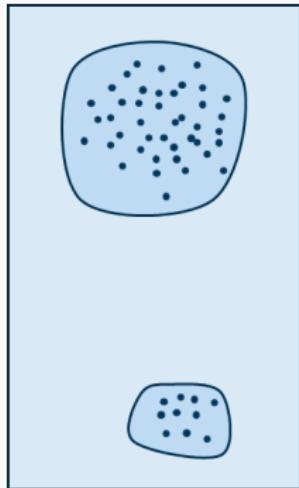
Node Table ▾

The screenshot shows a software interface titled "Node Table". At the top, there is a toolbar with icons for settings, refresh, search, and file operations. Below the toolbar is a header row containing the column names: "shared name", "name", "AverageShortestPathLength", "BetweennessCentrality", "ClosenessCentrality", and "ClusteringCoefficient". The main body of the table lists 12 rows of data, each representing a node. The data is as follows:

| shared name | name | AverageShortestPathLength | BetweennessCentrality | ClosenessCentrality | ClusteringCoefficient |
|-------------|----------|---------------------------|-----------------------|---------------------|-----------------------|
| YDL194W | YDL194W | 13.11693548 | 0.0 | 0.07623732 | 0.0 |
| YDR277C | YDR277C | 12.12096774 | 0.00806452 | 0.08250166 | 0.0 |
| YBR043C | YBR043C | 1.5 | 0.0 | 0.66666667 | 0.0 |
| YPR145W | YPR145W | 9.7983871 | 0.0 | 0.10205761 | 0.0 |
| YER054C | YER054C | 8.81854839 | 0.04427321 | 0.11339735 | 0.0 |
| YBR045C | YBR045C | 9.64516129 | 0.08528144 | 0.10367893 | 0.0 |
| YBL079W | YBL079W | 2.0 | 0.0 | 0.5 | 0.0 |
| YLR345W | YLR345W | 9.11693548 | 0.0 | 0.10968598 | 0.0 |
| YIL052C | YIL052C | 9.14919355 | 0.0 | 0.10929925 | 0.0 |
| YER056CA | YER056CA | 9.14919355 | 0.0 | 0.10929925 | 0.0 |

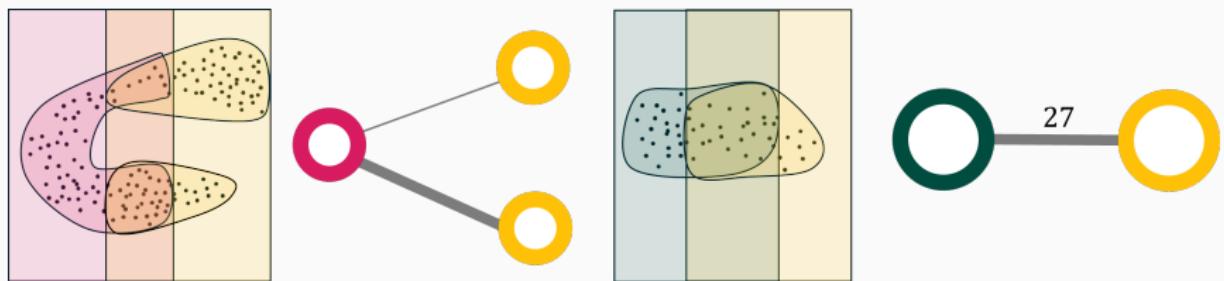
Styling Mapper: Vertices

- Node size/label → cluster size
- Node border color → associated level set/filter value/ball
- Node fill color → cluster dispersion



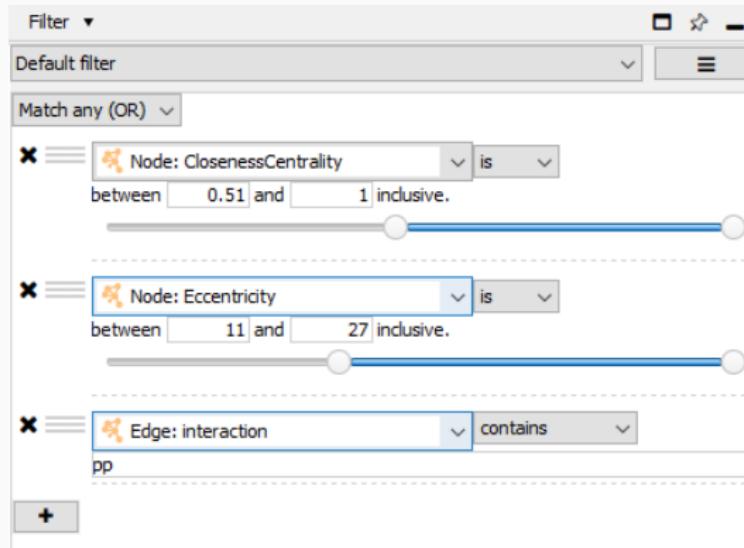
Styling Mapper: Edges

- Edge thickness-opacity → cluster intersection strength
- Edge label → cluster intersection size



Exploring the Graph

- Cytoscape can calculate classical network statistics
 - Centrality measures
 - Clustering coefficients
 - Modularity classes
- We can filter out nodes/edges by characteristics



Possible Capabilities

- Cytoscape is open source and was designed to be modified
- Possible projects here include:
 - Assign energy function to edges and apply layout algorithm
 - Animation between networks (say, from RBM to BM or reverse)
 - More graph algorithms (finding cores, clique detection, etc)

Section Map (!)

Why TDA?

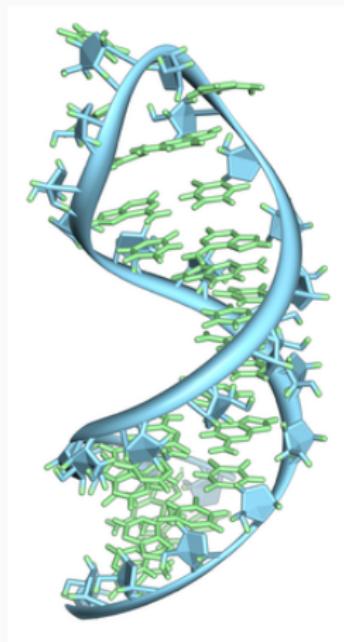
Mapper and Its Flavors

Cytoscape to the Rescue

Aptamers

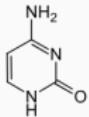
Mapping out Future Directions

- RNA (ribonucleic acid) and DNA (deoxyribonucleic acid) are polymers which carry genetic sequences and have additional structure



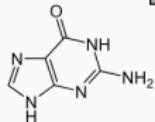
RNA and DNA

Cytosine



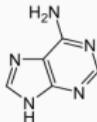
C

Guanine



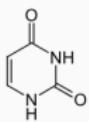
G

Adenine



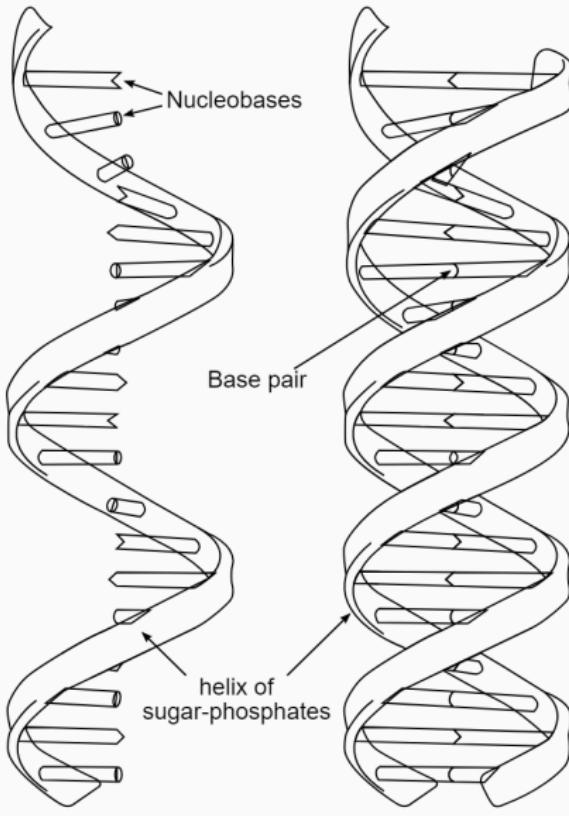
A

Uracil



U

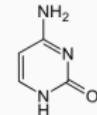
Nucleobases
of RNA



RNA

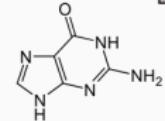
DNA

Cytosine



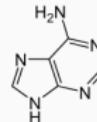
C

Guanine



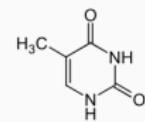
G

Adenine



A

Thymine

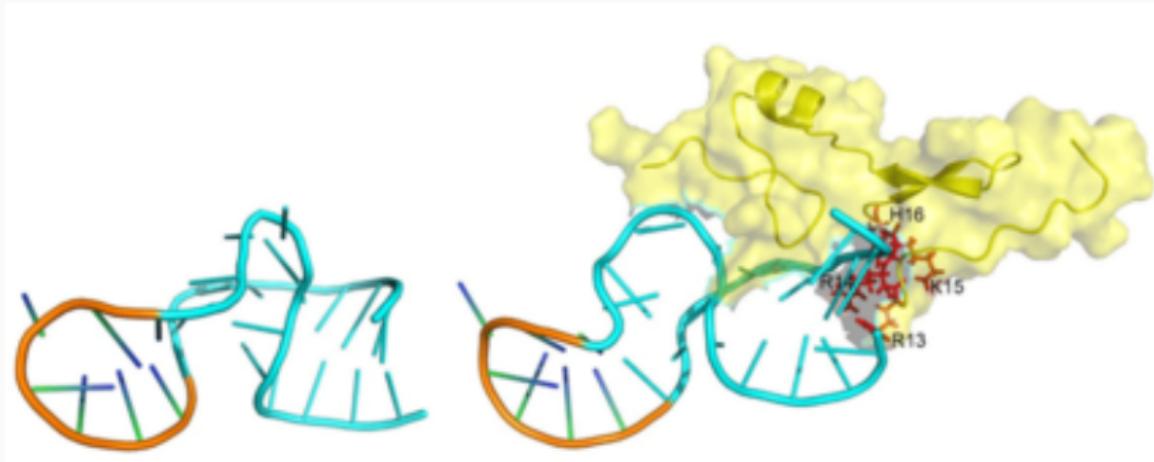


T

Nucleobases
of DNA

What Is an Aptamer?

- Aptamers are synthetic RNA molecules that bind to a specific target
- Similar function to antibodies, but much smaller
- **Genetic code not expressed**



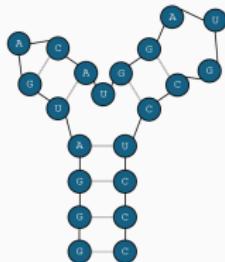
Comparing Aptamers

- For TDA to work we need a metric² by which to compare aptamers
- Aptamers have two characteristics: their genetic code and their structure
- Distance between sequences: Levenshtein distance
- Distance between structures: tree distance

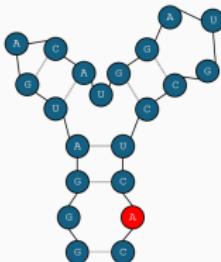
²ehhh...

Aptamer Metrics

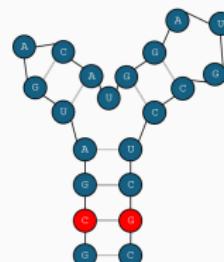
Aptamer X



Aptamer Y



Aptamer Z



GGGAUGACAUGGAUGGCCUCCC
((((((.)).((...))))))

δ_{Lev} : Levenshtein distance

δ_{Tree} : Tree distance

GGGAUGACAUGGAUGGCCUCAC
(.((((.)).((...)))).)

$$\delta_{\text{Lev}}(X, Y) = 1$$

$$\delta_{\text{Tree}}(X, Y) = 1$$

GGGCUGACAUGGAUGGCCGCC
((((((.)).((...))))))

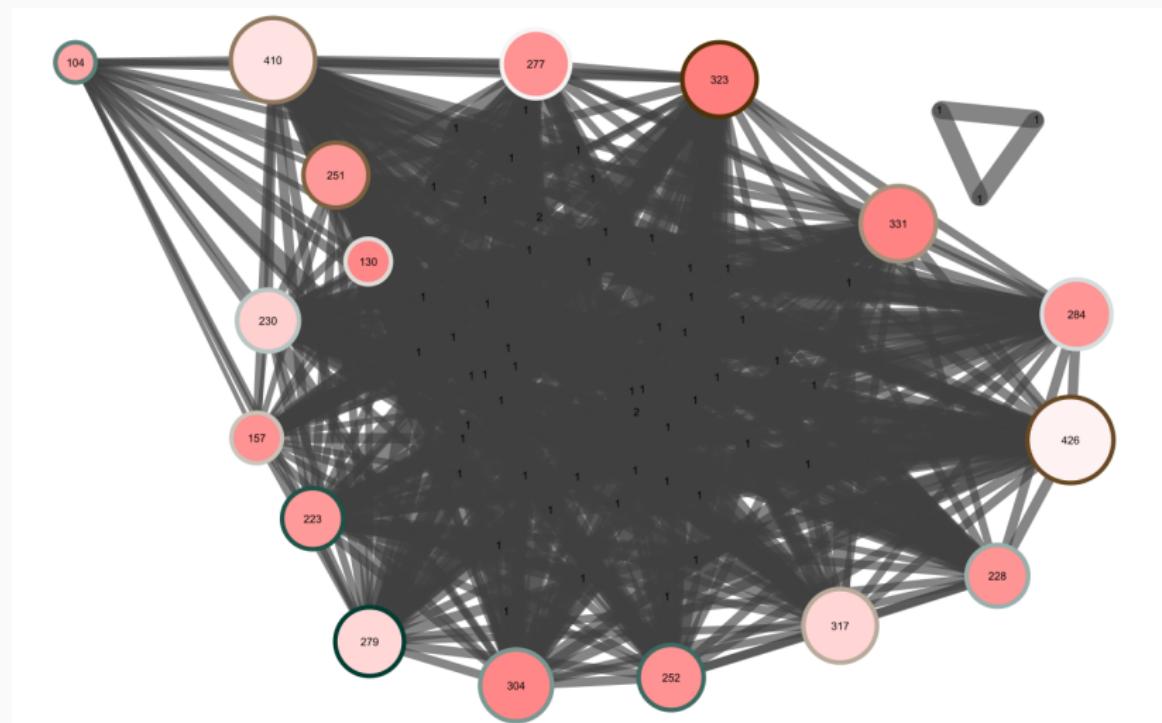
$$\delta_{\text{Lev}}(X, Z) = 2$$

$$\delta_{\text{Tree}}(X, Z) = 0$$

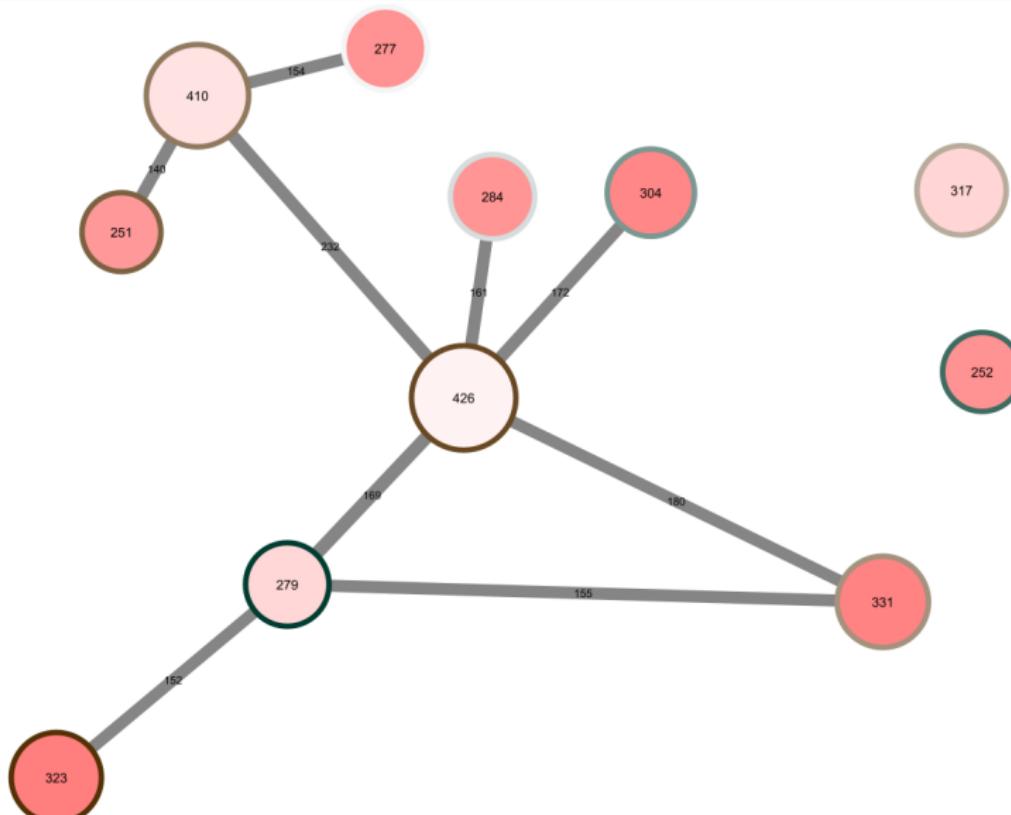
Aptamer Clustering With Mapper

- Flavor: Refined Ballmapper
- Idea: Ball using tree distance, cluster using Levenshtein distance
- Clustering method: single linkage hierarchical
- Vertices of the graph are clusters of aptamers related in both sequence and structure
- Graph structure may highlight families of aptamers or reveal other insights

Big Maptamer Graph



Pruned Maptamer Graph



Section Map (!)

Why TDA?

Mapper and Its Flavors

Cytoscape to the Rescue

Aptamers

Mapping out Future Directions

Clustering



Stability

Persistence

Return of Reeb

More Aptamers

