Mathematics of Data III: An Introduction to Topological Data Analysis

Yuan Yao

School of Mathematical Sciences
Peking University

July 13, 2011

1. Why Topological Methods?
   - Properties of Data Geometry
   - What Kind of Topological Methods?
2. Simplicial Complex for Data Representation
   - Simplicial Complex
3. Persistent Homology
   - Betti Number at Different Scales
   - Algebraic Theory
4. Some Applications
   - Coverage
   - Image
   - Molecular Dynamics
   - Progression Analysis of Disease

Introduction

- General method of manifold learning takes the following Spectral Kernal Embedding approach
  - construct a neighborhood graph of data, G
  - construct a positive semi-definite kernel on graphs, K
  - find global embedding coordinates of data by eigen-decomposition of $K = YY^T$
- Graph G may or may not reflect natural metric (e.g. similarity in genomics)
- Sometimes global embedding coordinates are not a good way to organize/visualize the data (e.g. $d > 3$)
- Sometimes all that is required is a qualitative view

Properties of Data Geometry

Fact

We Don't Trust Large Distances!

- In life or social sciences, distance (metric) are constructed using a notion of similarity (proximity), but have no theoretical backing (e.g. distance between faces, gene expression profiles, Jukes-Cantor distance between sequences)
- Small distances still represent similarity (proximity), but long distance comparisons hardly make sense

Properties of Data Geometry

Fact

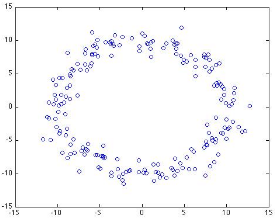
We Only Trust Small Distances a Bit!

- Both pairs are regarded as similar, but the strength of the similarity as encoded by the distance may not be so significant
- Similar objects lie in neighborhood of each other, which suffices to define topology

Properties of Data Geometry

Fact

Even Local Connections are Noisy, depending on observer's scale!



- Is it a circle, dots, or circle of circles?
- To see the circle, we ignore variations in small distance (tolerance for proximity)

So we need Topology here

- Distance measurements are noisy
- Physical device like human eyes may ignore differences in proximity (or as an average effect)
- Topology is the crudest way to capture invariants under distortions of distances
- At the presence of noise, one need topology varied with scales

Topology

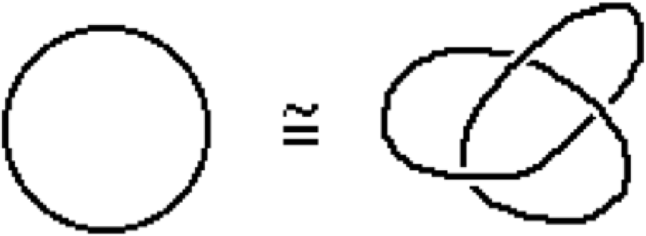


Figure: Homeomorphic

Topology

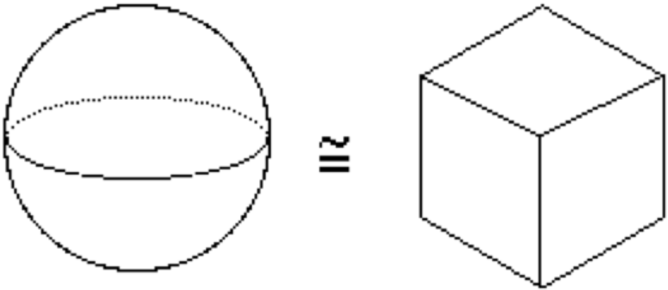


Figure: Homeomorphic

Topology

- The see that these pairs are *same* requires distortion of distances, i.e. stretching and shrinking
- We do not permit *tearing*, i.e. distorting distances in a discontinuous way
- How to make this precise, especially in discrete and noisy setting?

Topology

- We would like to say that all points within tolerance are the same
- Moreover, all non-zero distances beyond tolerance are the same, i.e. invariant under distortion

- Origins of Topology in Math
  - Leonhard Euler 1736, Seven Bridges of Königsberg
  - Johann Benedict Listing 1847, Vorstudien zur Topologie
  - J.B. Listing (orbituary) Nature 27:316-317, 1883. "qualitative geometry from the ordinary geometry in which quantitative relations chiefly are treated."

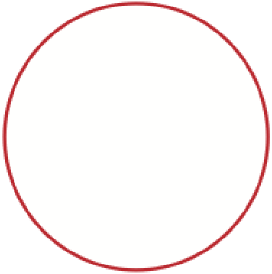
What kind of topology?

- Topology studies (global) mappings between spaces
- Point-set topology: continuous mappings on open sets
- Differential topology: differentiable mappings on smooth manifolds
  - Morse theory tells us topology of continuous space can be learned by discrete information on critical points
- Algebraic topology: homomorphisms on algebraic structures, the most concise encoder for topology
- Combinatorial topology: mappings on simplicial (cell) complexes
  - simplicial complex may be constructed from data
  - Algebraic, differential structures can be defined here

Topological Data Analysis

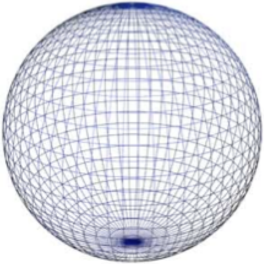
- What kind of topological information often useful
  - 0-homology: clustering or connected components
  - 1-homology: coverage of sensor networks; paths in robotic planning
  - 1-homology as obstructions: inconsistency in statistical ranking; harmonic flow games
  - high-order homology: high-order connectivity?
- How to compute homology in a stable way?
  - simplicial complexes for data representation
  - filtration on simplicial complexes
  - persistent homology

Betti Numbers



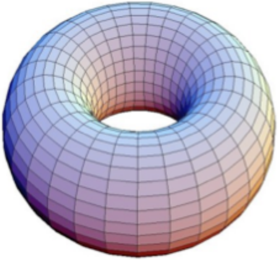
$\beta_0 = 1$, $\beta_1 = 1$, and $\beta_i = 0$ for $i \geq 2$

Betti Numbers



$\beta_0 = 1$, $\beta_1 = 0$, $\beta_2 = 0$, and $\beta_k = 0$ for $k \geq 3$

Betti Numbers



$$\beta_0 = 1,\ \beta_1 = 2,\ \beta_2 = 1, \text{ and } \beta_k = 0 \text{ for } k \geq 3$$

Simplicial Complexes for Data Representation

Definition (Simplicial Complex)

An abstract simplicial complex is a collection Σ of subsets of V which is closed under inclusion (or deletion), i.e. $\tau \in \Sigma$ and $\sigma \subseteq \tau$, then $\sigma \in \Sigma$.

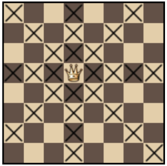
- Chess-board Complex
- Point cloud data:
  - Nerve complex
  - Cech, Rips, Witness complex
  - Mayer-Vietoris Blowup
- Term-document cooccurance complex
- Clique complex in pairwise comparison graphs
- Strategic complex in flow games

Chess-board Complex

Definition (Chess-board Complex)

Let V be the positions on a Chess board. Σ collects position subsets of V where one can place queens (rooks) without capturing each other.

- Closedness under deletion: if $\sigma \in \Sigma$ is a set of “safe” positions, then any subset $\tau \subseteq \sigma$ is also a set of “safe” positions



Nerve complex

Definition (Nerve Complex)

Define a cover of X, $X = \cup_\alpha U_\alpha$. $V = \{U_\alpha\}$ and define $\Sigma = \{U_I : \cap_{\alpha \in I} U_I \neq \emptyset\}$.

- Closedness under deletion
- Can be applied to any topological space X
- In a metric space (X, d), if $U_\alpha = B_\epsilon(t_\alpha) := \{x \in X : d(x - t_\alpha) \leq \epsilon\}$, we have Čech complex C_ϵ.
- Nerve Theorem: if every U_I is contractible, then X has the same homotopy type as Σ.

Example: Nerve/Čech Complex

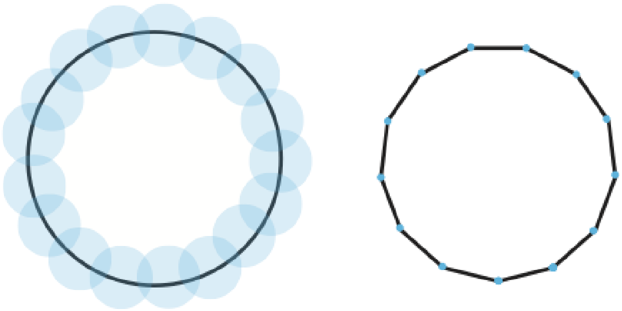


Figure: Čech complex of a circle, C_ϵ, covered by a set of balls.

Vietoris-Rips complex

- Čech complex is hard to compute, even in Euclidean space
- One can easily compute an upper bound for Čech complex
  - Construct a Čech subcomplex of 1-dimension, i.e. a graph with edges connecting point pairs whose distance is no more than ϵ.
  - Find the clique complex, i.e. maximal complex whose 1-skeleton is the graph above, where every k-clique is regarded as a $k-1$ simplex

Definition (Vietoris-Rips Complex)

Let $V = \{x_\alpha \in X\}$. Define
$VR_\epsilon = \{U_I \subseteq V : d(x_\alpha, x_\beta) \le \epsilon, \alpha, \beta \in I\}$.

Example: Rips Complex

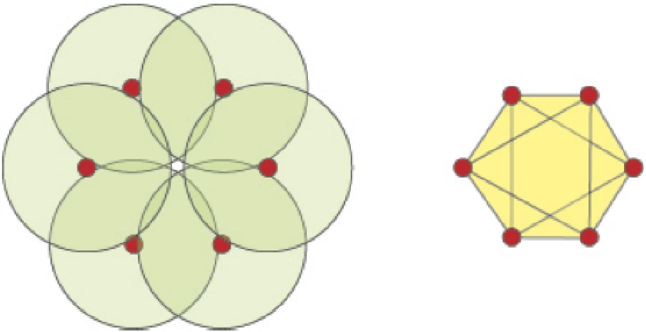


Figure: Left: Čech complex gives a circle; Right: Rips complex gives a sphere S^2.

Generalized Vietoris-Rips for Symmetric Relations

Definition (Symmetric Relation Complex)

Let V be a set and a symmetric relation $R = \{(u, v)\} \subseteq V^2$ such that $(u, v) \in R \Rightarrow (v, u) \in R$. Σ collects subsets of V which are in pairwise relations.

- Closedness under deletion: if $\sigma \in \Sigma$ is a set of related items, then any subset $\tau \subseteq \sigma$ is a set of related items
- Generalized Vietoris-Rips complex beyond metric spaces
- E.g. Zeeman's tolerance space
- C.H. Dowker defines simplicial complex for unsymmetric relations

Sandwich Theorems

- Rips is easier to compute than Cech
  - even so, Rips is exponential to dimension generally
- However Vietoris-Rips CAN NOT preserve the homotopy type as Cech
- But there is still a hope to find a lower bound on homology –

Theorem ("Sandwich")

$$VR_{\epsilon} \subseteq C_{\epsilon} \subseteq VR_{2\epsilon}$$

- If a homology group "persists" through $R_{\epsilon} \to R_{2\epsilon}$, then it must exists in C_{ϵ}; but not the vice versa.

A further simplification: Witness complex

Definition (Strong Witness Complex)

Let $V = \{t_\alpha \in X\}$. Define
$W_\epsilon^s = \{U_I \subseteq V : \exists x \in X, \forall \alpha \in I, d(x, t_\alpha) \leq d(x, V) + \epsilon\}$.

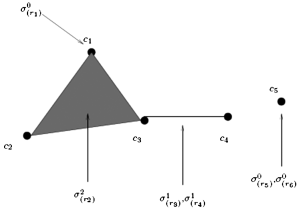
Definition (Week Witness Complex)

Let $V = \{t_\alpha \in X\}$. Define
$W_\epsilon^w = \{U_I \subseteq V : \exists x \in X, \forall \alpha \in I, d(x, t_\alpha) \leq d(x, V_{-I}) + \epsilon\}$.

- V can be a set of landmarks, much smaller than X
- Monotonicity: $W_\epsilon^* \subseteq W_{\epsilon'}^*$ if $\epsilon \leq \epsilon'$
- But not easy to control homotopy types between W^* and X

Term-Document Co-occurrence Complex

	c_1	c_2	c_3	c_4	c_5
r_1	1	0	0	0	0
r_2	1	1	1	0	0
r_3	0	0	1	1	0
r_4	0	0	1	1	0
r_5	0	0	0	0	1
r_6	0	0	0	0	1



- Left is a term-document co-occurrence matrix
- Right is a simplicial complex representation of terms
- Connectivity analysis captures more information than Latent Semantic Index (Li & Kwong 2009)

Strategic Simplicial Complex for Flow Games

	O	F
O	3, 2	0, 0
F	0, 0	2, 3

(a) Battle of the sexes

$$\begin{tikzcd}
(O,O) & (O,F) \arrow[l, "2"'] \arrow[d, "2"] \\
(F,O) \arrow[u, "3"'] \arrow[r, "3"] & (F,F)
\end{tikzcd}$$

- Strategic simplicial complex is the clique complex of pairwise comparison graph above, inspired by ranking
- Every game can be decomposed as the direct sum of potential games and zero-sum games (harmonic games) (Candogan, Menache, Ozdaglar and Parrilo 2010)

Example I: Persistent Homology of Čech Complexes

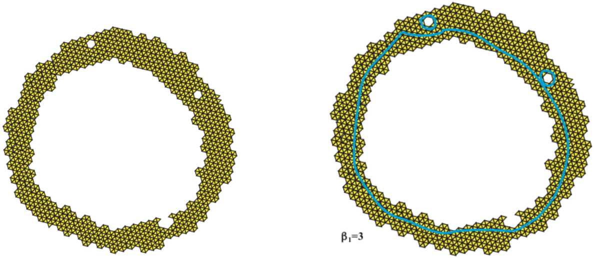


Figure: Scale ϵ_1: $\beta_0 = 1$, $\beta_1 = 3$

Example I: Persistent Homology of Čech Complexes

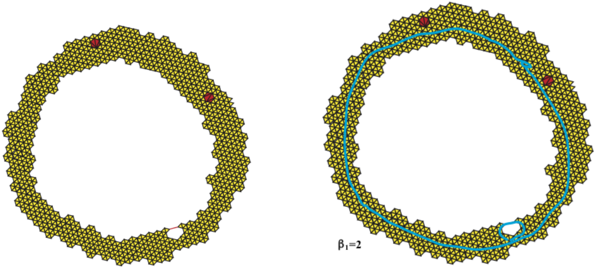


Figure: Scale ϵ_1: $\beta_0 = 1$, $\beta_1 = 2$

Example II: Persistence 0-Homology induced by Height Function

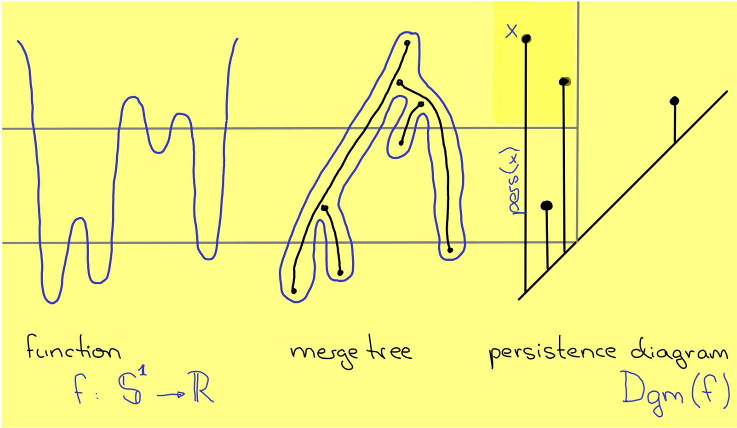
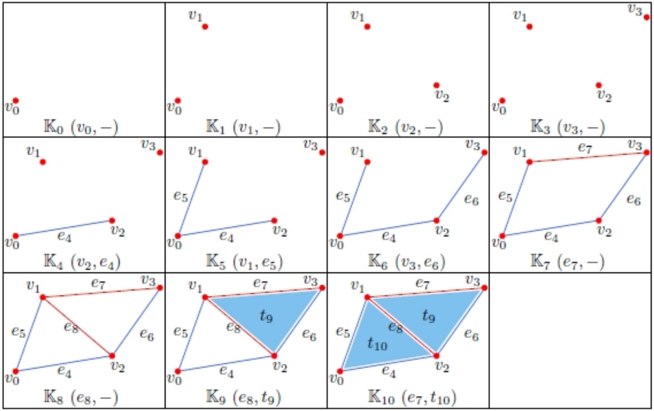
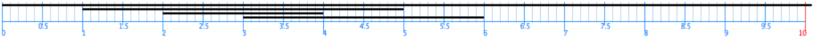


Figure: The birth and death of connected components.

Example III: Persistent Homology as Online Algorithm to Track Topology Changements



Persistent Betti Numbers: Barcodes



Barcodes: Dimension 0

Barcodes: Dimension 1

- Toolbox: JPlex (http://comptop.stanford.edu/)
  - Java version of Plex, work with matlab
  - Rips, Witness complex, Persistence Homology
- Other Choices: Plex 2.5 for Matlab (not maintained any more), Dionysus (Dimitry Morozov)

Persistent Homology: Algebraic Theory [Zormorodian-Carlsson]

- All above gives rise to a filtration of simplicial complex

$$\emptyset = \Sigma_0 \subseteq \Sigma_1 \subseteq \Sigma_2 \subseteq \ldots$$

- Functoriality of inclusion: there are homomorphisms between homology groups

$$0 \to H_1 \to H_2 \to \ldots$$

- A persistent homology is the image of H_i in H_j with $j > i$.

Persistent 0-Homology of Rips Complex

- Equivalent to single-linkage clustering
- Barcode is the single linkage dendrogram (tree) without labels
- Kleinberg's Impossibility Theorem for clustering: no clustering algorithm satisfies scale invariance, richness, and consistency
- Memoli & Carlsson 2009: single-linkage is the unique persistent clustering with scale invariance
- Open: but, is persistence the necessity for clustering?
- Notes: try matlab command `linkage` for single-linkage clustering.

Application I: Sensor Network Coverage by Persistent Homology

- V. de Silva and R. Ghrist (2005) Coverage in sensor networks via persistent homology.
- Ideally sensor communication can be modeled by Rips complex
  - two sensors has distance within a short range, then two sensors receive strong signals;
  - two sensors has distance within a middle range, then two sensors receive weak signals;
  - otherwise no signals

Sandwich Theorem

Theorem (de Silva-Ghrist 2005)

Let X be a set of points in R^d and $C_\epsilon(X)$ the Čech complex of the cover of X by balls of radius $\epsilon/2$. Then there is chain of inclusions

$$R_{\epsilon'}(X) \subset C_\epsilon(X) \subset R_\epsilon(X) \quad \textit{whenever} \quad \frac{\epsilon}{\epsilon'} \geq \sqrt{\frac{2d}{d+1}}.$$

Moreover, this ratio is the smallest for which the inclusions hold in general.

Note: this gives a sufficient condition to detect holes in sensor network coverage

- Čech complex is hard to compute while Rips is easy;
- If a hole persists from $R_{\epsilon'}$ to R_ϵ, then it must exists in C_ϵ.

Persistent 1-Homology in Rips Complexes

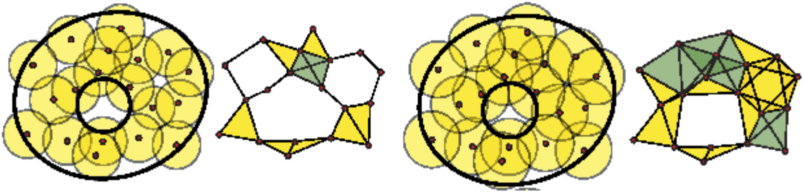


Figure: Left: $R_{\epsilon'}$; Right: R_{ϵ}. The middle hole persists from $R_{\epsilon'}$ to R_{ϵ}.

Application II: Natural Image Statistics

- G. Carlsson, V. de Silva, T. Ishkanov, A. Zomorodian (2008) On the local behavior of spaces of natural images, *International Journal of Computer Vision*, 76(1):1-12.
- An image taken by black and white digital camera can be viewed as a vector, with one coordinate for each pixel
- Each pixel has a "gray scale" value, can be thought of as a real number (in reality, takes one of 255 values)
- Typical camera uses tens of thousands of pixels, so images lie in a very high dimensional space, call it pixel space, $\mathcal{P}$

Natural Image Statistics

- **D. Mumford**: What can be said about the set of images $\mathcal{I} \subseteq \mathcal{P}$ one obtains when one takes many images with a digital camera?
- **Lee, Mumford, Pedersen**: Useful to study local structure of images statistically

Natural Image Statistics

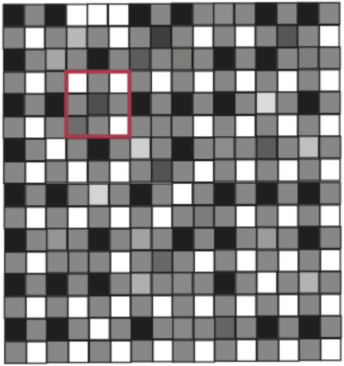


Figure: 3×3 patches in images

Natural Image Statistics

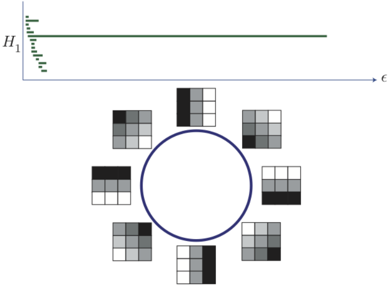
Lee-Mumford-Pedersen [LMP] study only high contrast patches.

- Collect: 4.5M high contrast patches from a collection of images obtained by van Hateren and van der Schaaf
- Normalize mean intensity by subtracting mean from each pixel value to obtain patches with mean intensity $= 0$
- Puts data on an 8-D hyperplane, $\approx R^8$
- Furthermore, normalize contrast by dividing by the norm, so obtain patches with norm $= 1$, whence data lies on a 7-D ellipsoid, $\approx S^7$

Natural Image Statistics: Primary Circle

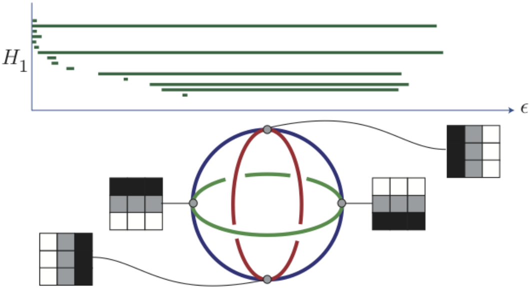
High density subsets $\mathcal{M}(k = 300, t = 0.25)$:

- Codensity filter: $d_k(x)$ be the distance from x to its k-th nearest neighbor
  - the lower $d_k(x)$, the higher density of x
- Take $k = 300$, the extract $5,000$ top $t = 25\%$ densest points, which concentrate on a primary circle



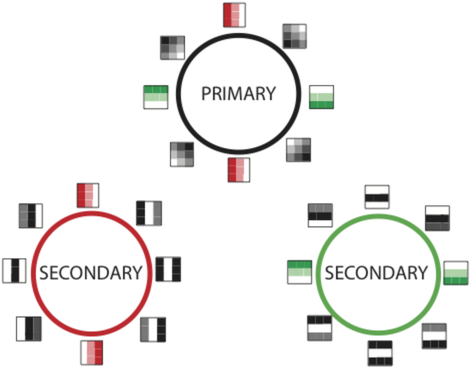
Natural Image Statistics: Three Circles

- Take $k = 15$, the extract $5,000$ top 25% densest points, which shows persistent $\beta_1 = 5$, 3-circle model

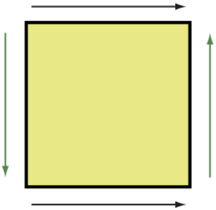


Natural Image Statistics: Three Circles

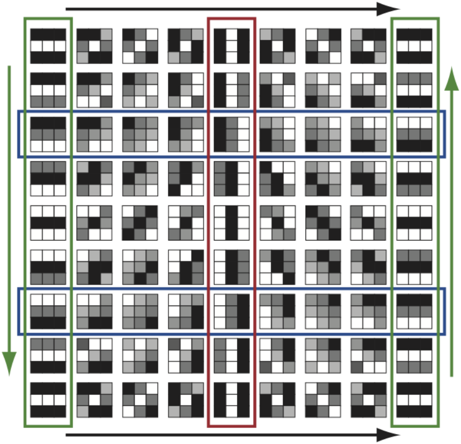
Generators for 3 circles



Natural Image Statistics: Klein Bottle



Natural Image Statistics: Klein Bottle Model



Application III: Persistent Homology and Discrete Morse Theory

- Persistent homology gives a pairing (birth-death) between a simplex and its co-dimensional one faces
- It leads to a particular implementation of Robin Forman's combinatorial gradient field
- Thus Persistent homology is equivalent to discrete Morse Theory by Robin Forman

Morse Theory and Reeb graph

- a nice (Morse) function: $h : \mathcal{X} \to \mathbb{R}$, on a smooth manifold $\mathcal{X}$
- topology of $\mathcal{X}$ reconstructed from level sets $h^{-1}(t)$
- topological of $h^{-1}(t)$ only changes at 'critical values'
- Reeb graph: a simplified version, contracting into points the connected components in $h^{-1}(t)$

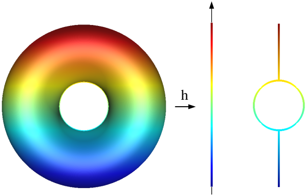
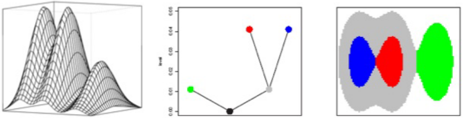


Figure: Construction of Reeb graph; h maps each point on torus to its height.

In applications.

Reeb graph has found various applications in computational geometry, statistics under different names.

- computer science: contour trees, reeb graphs
- statistics: density cluster trees (Hartigan)



Mapper: an extension for topological data analysis

[*Singh-Memoli-Carlsson. Eurograph-PBG, 2007*] Given a data set $\mathcal{X}$,

- choose a filter map $h : \mathcal{X} \to T$, where T is a topological space such as $\mathbb{R}$, S^1, $\mathbb{R}^d$, etc.
- choose a cover $T \subseteq \cup_\alpha U_\alpha$
- cluster/partite level sets $h^{-1}(U_\alpha)$ into $V_{\alpha,\beta}$
- graph representation: a node for each $V_{\alpha,\beta}$, an edge between $(V_{\alpha_1,\beta_1}, V_{\alpha_2,\beta_2})$ iff $U_{\alpha_1} \cap U_{\alpha_2} \neq \emptyset$ and $V_{\alpha_1,\beta_1} \cap V_{\alpha_2,\beta_2} \neq \emptyset$.
- extendable to simplicial complex representation.

Note: it extends Morse theory from $\mathbb{R}$ to general topological space T; may lead to a particular implementation of Nerve theorem through filter map h.

An example with real valued filter

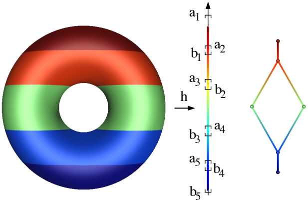


Figure: An illustration of Mapper.

Note:

- degree-one nodes contain local minima/maxima;
- degree-three nodes contain saddle points (critical points);
- degree-two nodes consist of regular points

Example: RNA Tetraloop

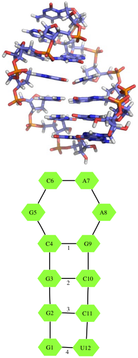


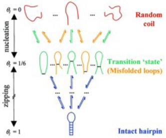
Figure: RNA GCAA-Tetraloop

Biological relevance:

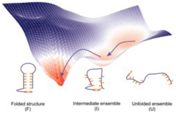
- serve as nucleation site for RNA folding
- form sequence specific tertiary interactions
- protein recognition sites
- certain Tetraloops can pause RNA transcription

Note: simple, but, biological debates over intermediate states on folding pathways

Debates: Two-state vs. Multi-state Models



(a) 2-state model



(b) multi-state model

- 2-state: transition state with any one stem base pair, from thermodynamic experiments [*Ansari A, et al. PNAS, 2001, 98: 7771-7776*]
- multi-state: there is a stable intermediate state, which contains collapsed structures, from kinetic measurements [*Ma H, et al. PNAS, 2007, 104:712-6*]
- experiments: no structural information
- computer simulations at full-atom resolution:
  - exisitence of intermediate states
  - if yes, what's the structure?

Mapper with density filters in biomolecular folding

Reference: Bowman-Huang-Yao et al. J. Am. Chem. Soc. 2008; Yao, Sun, Huang, et al. J. Chem. Phys. 2009.

- densest regions (energy basins) may correspond to metastates (e.g. folded, extended)
- intermediate/transition states on pathways connecting them are relatively sparse

Therefore with Mapper

- clustering on density level sets helps separate and identify metastates and intermediate/transition states
- graph representation reflects kinetic connectivity between states

A vanilla version

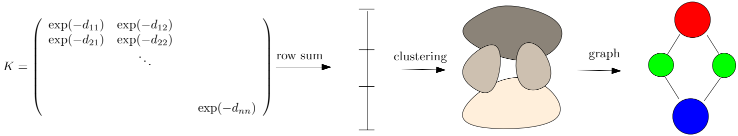


Figure: Mapper Flow Chart

1. Kernel density estimation $h(x) = \sum_i K(x, x_i)$ with Hamming distance for contact maps
2. Rank the data by h and divide the data into n overlapped sets
3. Single-linkage clustering on each level sets
4. Graphical representation

Mapper output for Unfolding Pathways

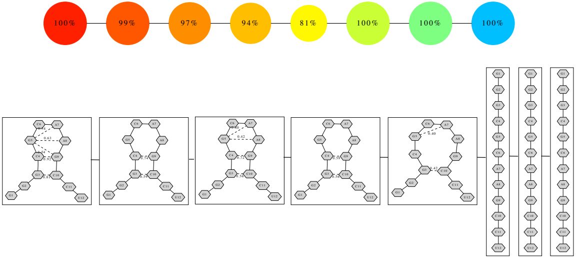


Figure: Unfolding pathway

Mapper output for Refolding Pathways

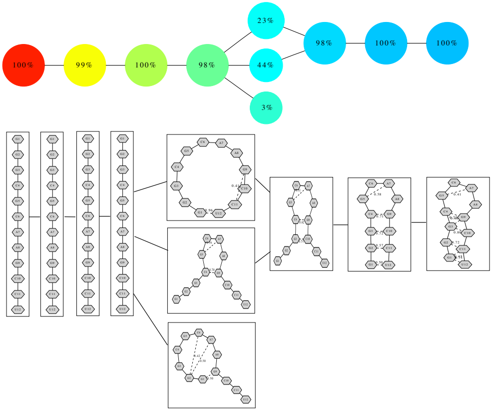
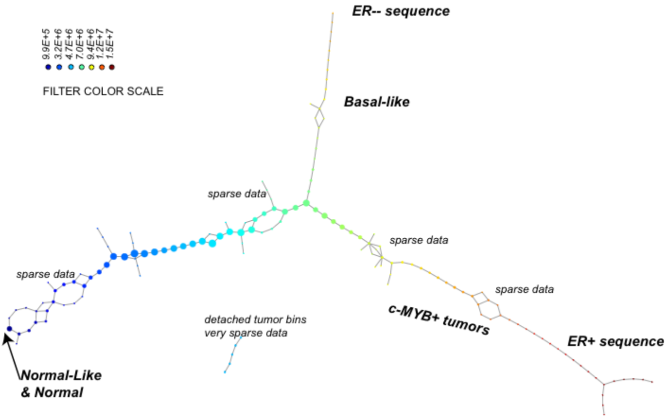


Figure: Refolding pathway

Application IV: Progression Analysis for Breast Cancer

- Nicolau, Levine, Carlsson, *PNAS*, 2010
- Deviation functions from normal tissues are used as filters (Morse-type functions)
- Mapper (Reeb Graph) with such filters leads to Progression Analysis of Disease

PAD analysis of the NKI data



Reference

- Edelsbrunner, Letscher, and Zomorodian (2002) Topological Persistence and Simplification.
- Ghrist, R. (2007) Barcdes: the Persistent Topology of Data. *Bulletin of AMS*, 45(1):61-75.
- Edelsbrunner, Harer (2008) Persistent Homology - a survey. *Contemporary Mathematics*.
- Carlsson, G. (2009) Topology and Data. *Bulletin of AMS*, 46(2):255-308.