An Introduction to Topological Data Analysis

Yuan Yao

Department of Mathematics
HKUST

November, 2017

Methods for Summarizing or Visualizing a Geometry

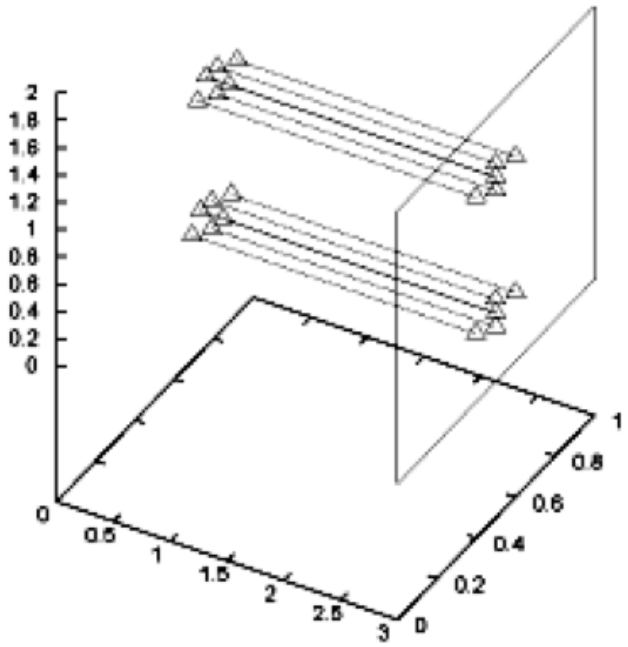


Figure: Linear projection (PCA, MDS, variable selection, etc)

Methods for Summarizing or Visualizing a Geometry

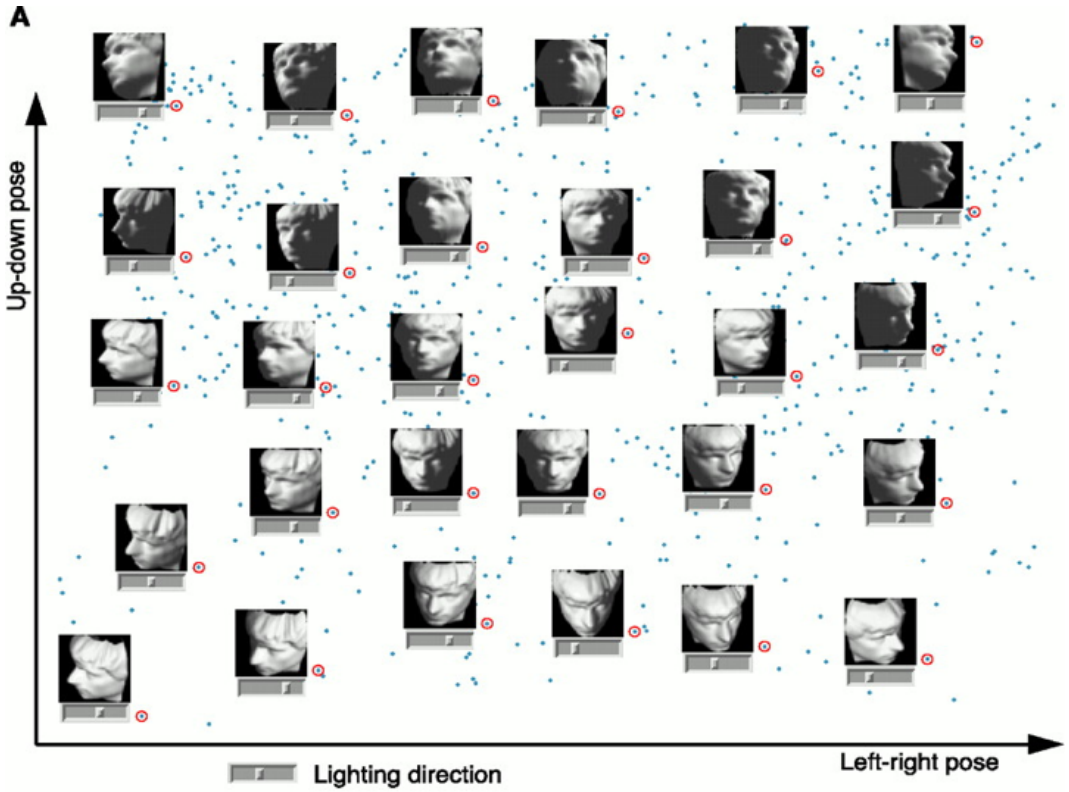


Figure: Nonlinear Dimensionality Reduction (ISOMAP, LLE etc.)

Geometric Data Reduction

- 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 $K = YY^T$
- Sometimes 'distance metric' is just a similarity measure (nonmetric MDS, ordinal embedding)
- Sometimes coordinates are not a good way to organize/visualize the data (e.g. $d > 3$)
- Sometimes all that is required is a qualitative view

Topology

- 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."

RNA hairpin folding pathways

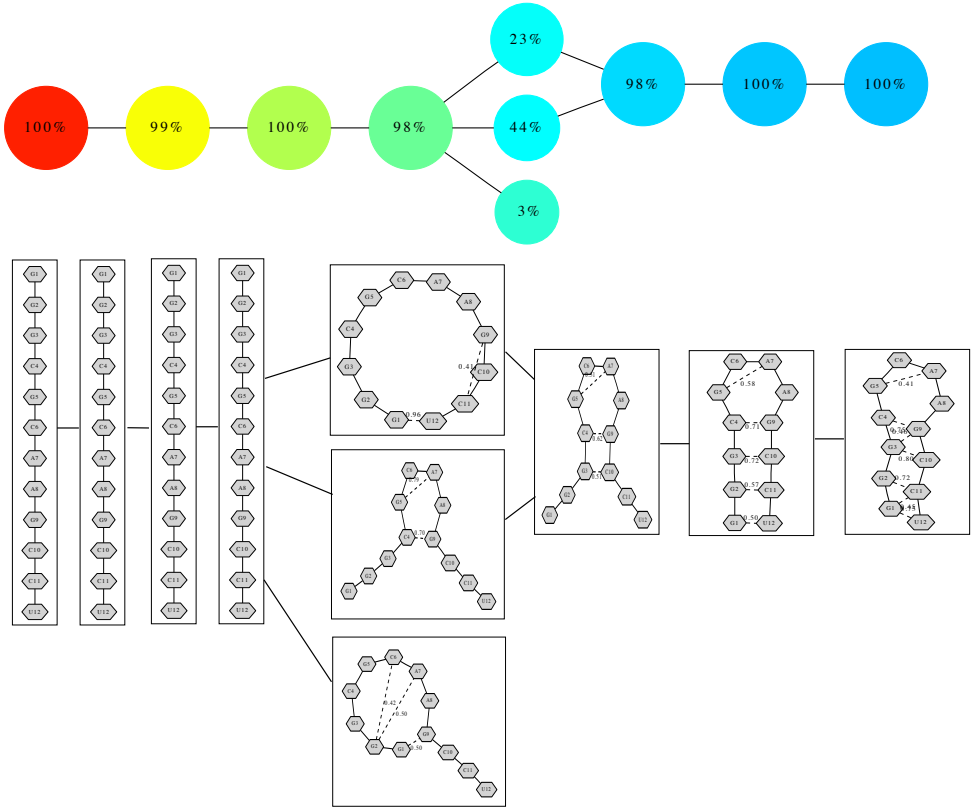


Figure: Jointly with Xuhui Huang, Jian Sun, Greg Bowman, Gunnar Carlsson, Leo Guibas, and Vijay Pande, *JACS'08*, *JCP'09*

Progression of Breast Cancer with gene expression profiles

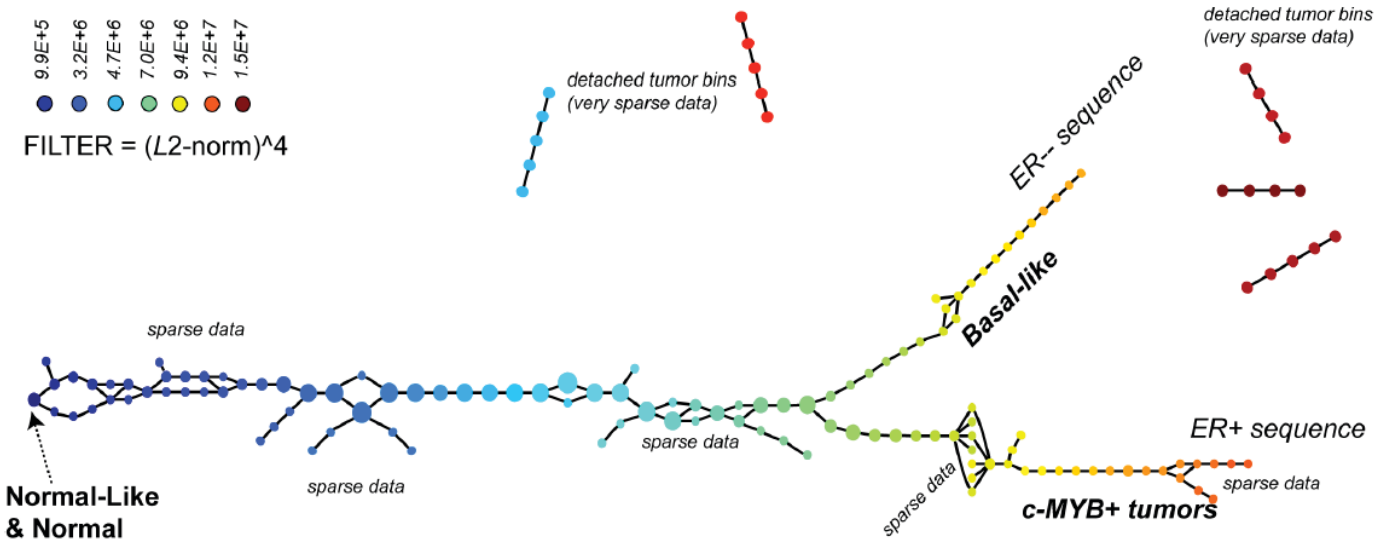


Figure: Monica Nicolau, A. Levine, and Gunnar Carlsson, *PNAS'10*

Key elements

- Coordinate free representation
- Invariance under deformations
- Compressed qualitative representation

Topology

- To see points in neighborhood the *same* requires distortion of distances, i.e. stretching and shrinking
- We do not permit *tearing*, i.e. distorting distances in a discontinuous way

Continous Topology

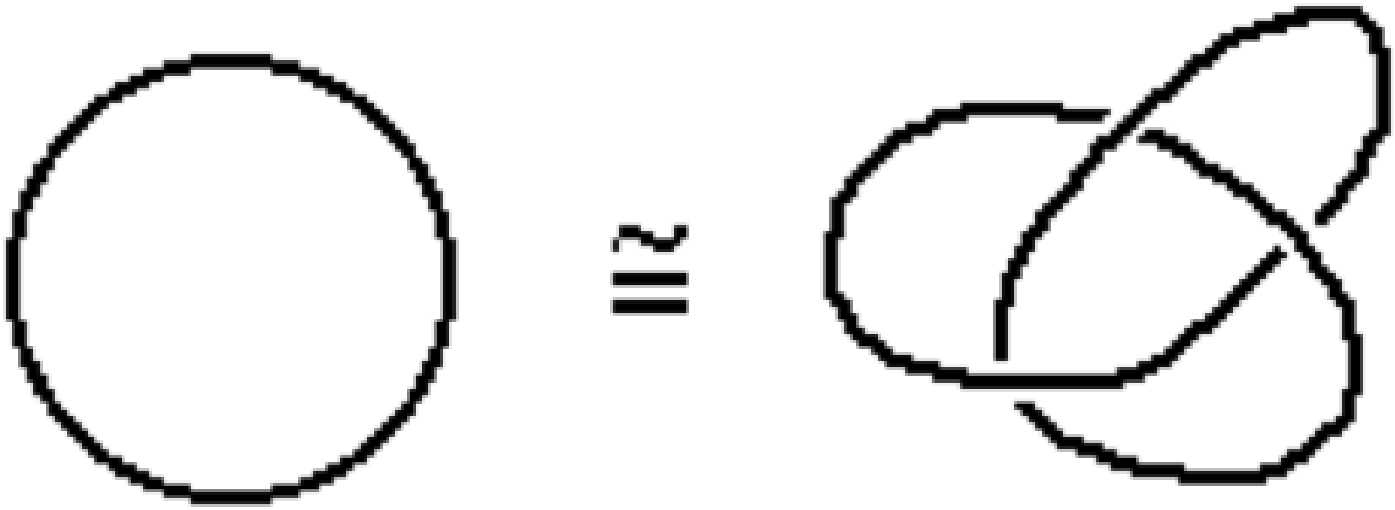


Figure: Homeomorphic

Continuous Topology

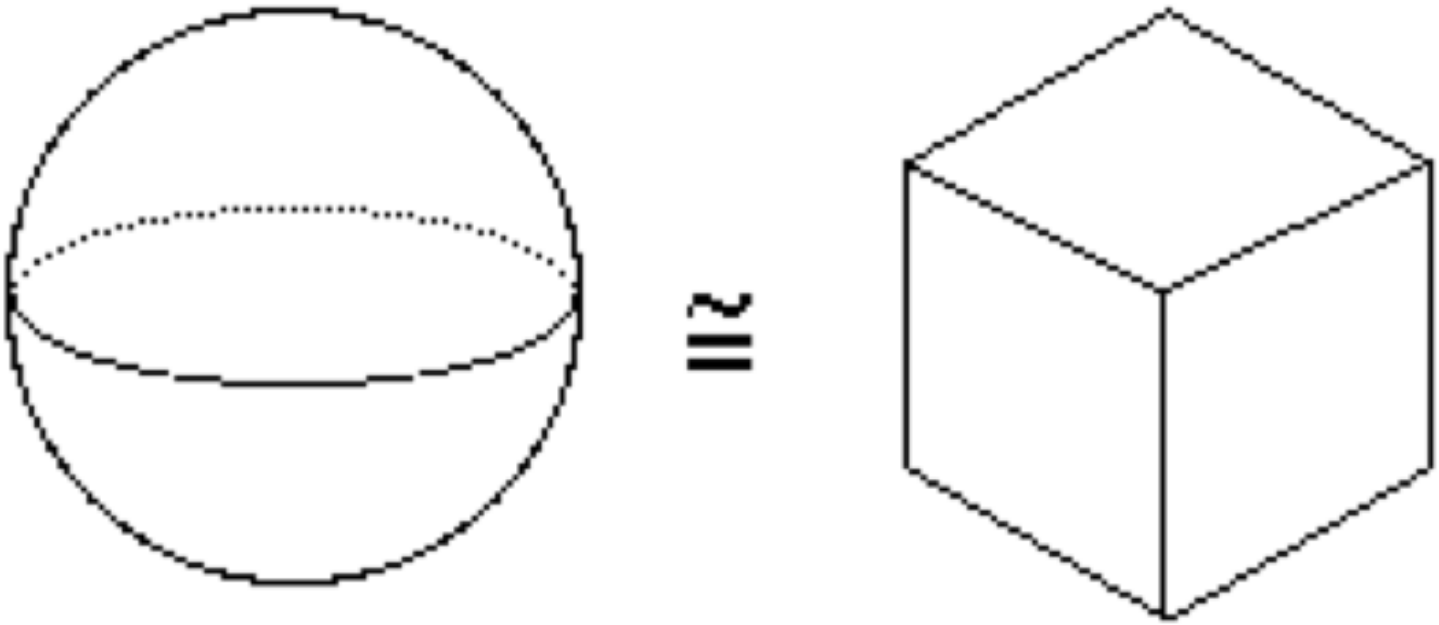


Figure: Homeomorphic

Discrete case?

How does topology make sense, in discrete and noisy setting?

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

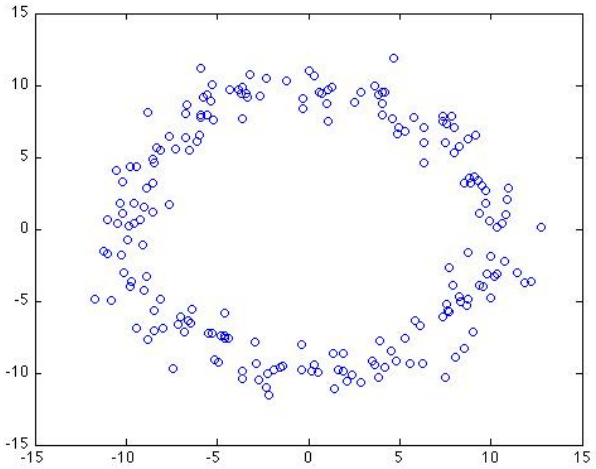
We Only Trust Small Distances a Bit!

- 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 for robustness against metric distortions

- 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

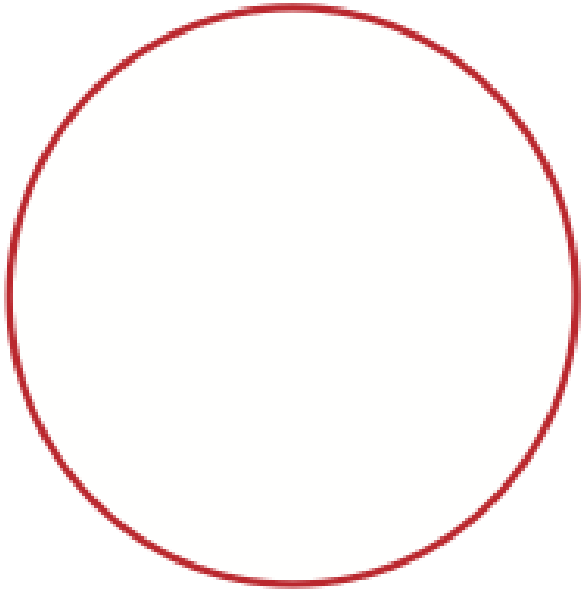
What kind of topology?

- 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: the number of i-dim holes



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

Betti Numbers: the number of i-dim holes

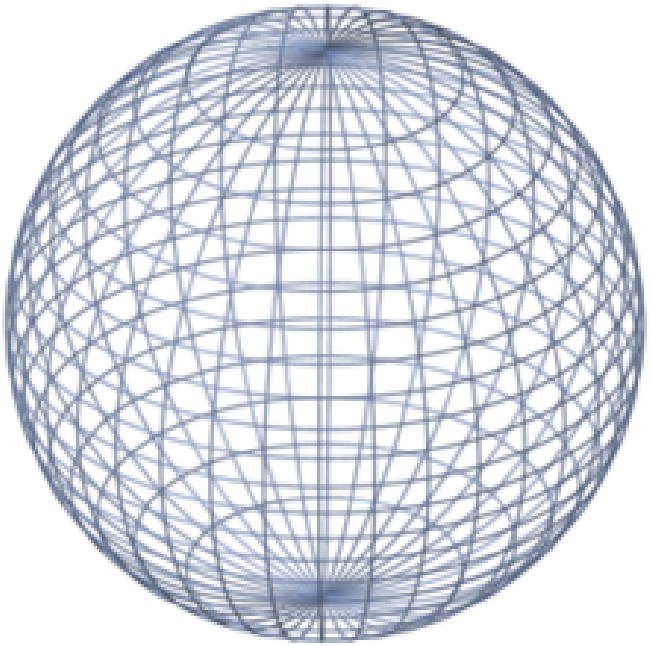
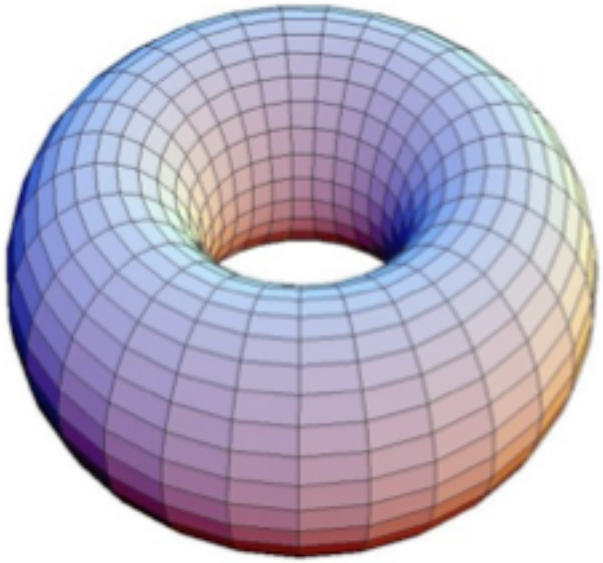


Figure: Sphere: $\beta_0 = 1$, $\beta_1 = 0$, $\beta_2 = 1$, and $\beta_k = 0$ for $k \geq 3$

Betti Numbers: the number of i-dim holes



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

Betti Numbers and Homology Groups

- Betti numbers are computed as dimensions of Boolean vector spaces (E. Noether, $\mathbb{Z}_2$-homology group)

Betti Numbers and Homology Groups

- Betti numbers are computed as dimensions of Boolean vector spaces (E. Noether, $\mathbb{Z}_2$-homology group)
- $\beta_i(X) = dim H_i(X, \mathbb{Z}_2)$, $\mathbb{Z}_2$-homology or more general Homology group associated with any fields or integral domain (e.g. $\mathbb{Z}$, $\mathbb{Q}$, and $\mathbb{R}$)

Betti Numbers and Homology Groups

- Betti numbers are computed as dimensions of Boolean vector spaces (E. Noether, $\mathbb{Z}_2$-homology group)
- $\beta_i(X) = dim H_i(X, \mathbb{Z}_2)$, $\mathbb{Z}_2$-homology or more general Homology group associated with any fields or integral domain (e.g. $\mathbb{Z}$, $\mathbb{Q}$, and $\mathbb{R}$)
- $H_i(X)$ is *functorial*, i.e. continuous mapping $f : X \to Y$ induces linear transformation $H_i(f) : H_i(X) \to H_i(Y)$, structure preserving

Betti Numbers and Homology Groups

- Betti numbers are computed as dimensions of Boolean vector spaces (E. Noether, $\mathbb{Z}_2$-homology group)
- $\beta_i(X) = dim H_i(X, \mathbb{Z}_2)$, $\mathbb{Z}_2$-homology or more general Homology group associated with any fields or integral domain (e.g. $\mathbb{Z}$, $\mathbb{Q}$, and $\mathbb{R}$)
- $H_i(X)$ is *functorial*, i.e. continuous mapping $f : X \to Y$ induces linear transformation $H_i(f) : H_i(X) \to H_i(Y)$, structure preserving
- computation is simple linear algebra over fields or integers

Betti Numbers and Homology Groups

- Betti numbers are computed as dimensions of Boolean vector spaces (E. Noether, $\mathbb{Z}_2$-homology group)
- $\beta_i(X) = dim H_i(X, \mathbb{Z}_2)$, $\mathbb{Z}_2$-homology or more general Homology group associated with any fields or integral domain (e.g. $\mathbb{Z}$, $\mathbb{Q}$, and $\mathbb{R}$)
- $H_i(X)$ is *functorial*, i.e. continuous mapping $f : X \to Y$ induces linear transformation $H_i(f) : H_i(X) \to H_i(Y)$, structure preserving
- computation is simple linear algebra over fields or integers
- data representation by *simplicial complexes*

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 $\sigma \in \Sigma$.

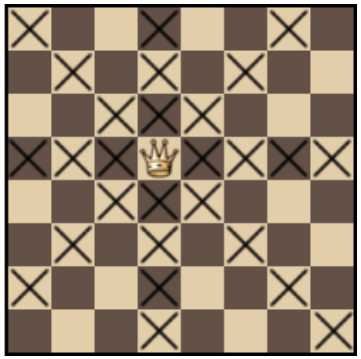
- Chess-board Complex
- Term-document cooccurance complex
- Nerve complex
- Point cloud data in metric spaces:
  - Čech, Rips, Witness complex
  - Mayer-Vietoris Blowup
- Clique complex in pairwise comparison graphs
- Strategic complex in game theory

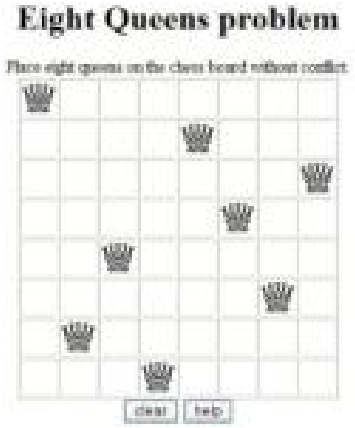
Chess-board Complex

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.

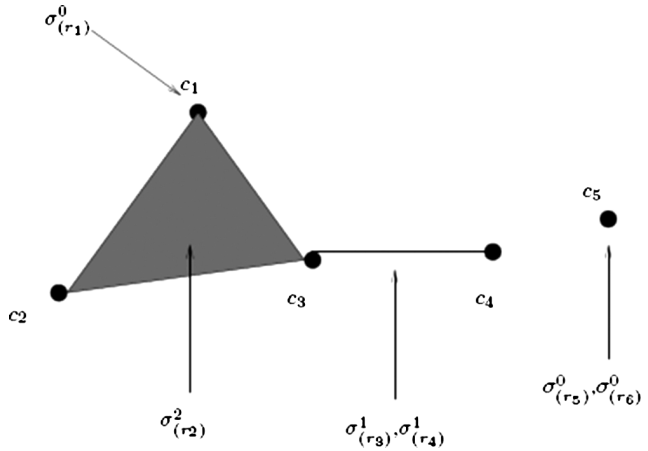
- Closedness under deletion: if $\sigma \in \Sigma$ is a set of "safe" positions, then any subset $\tau \subseteq \sigma$ is also a set of "safe" positions





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)

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
- Nerve Theorem: if every U_I is contractible, then X has the same homotopy type as Σ.

Nerve complex example

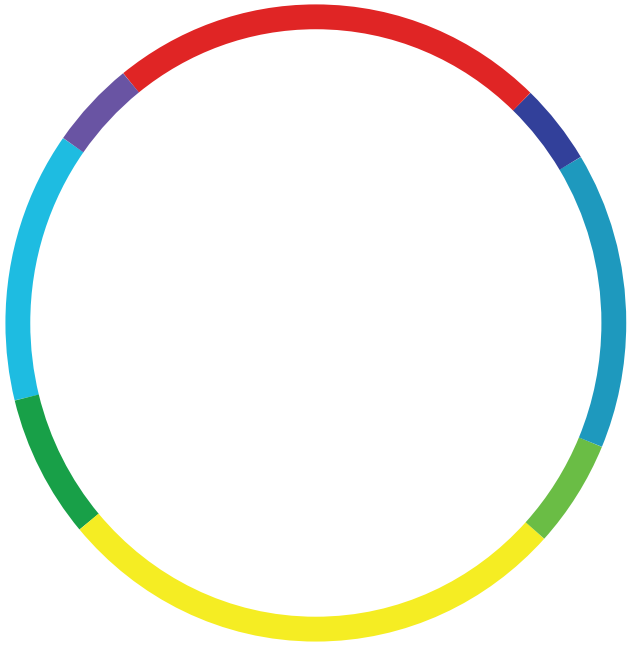


Figure: Covering of circle

Nerve complex example

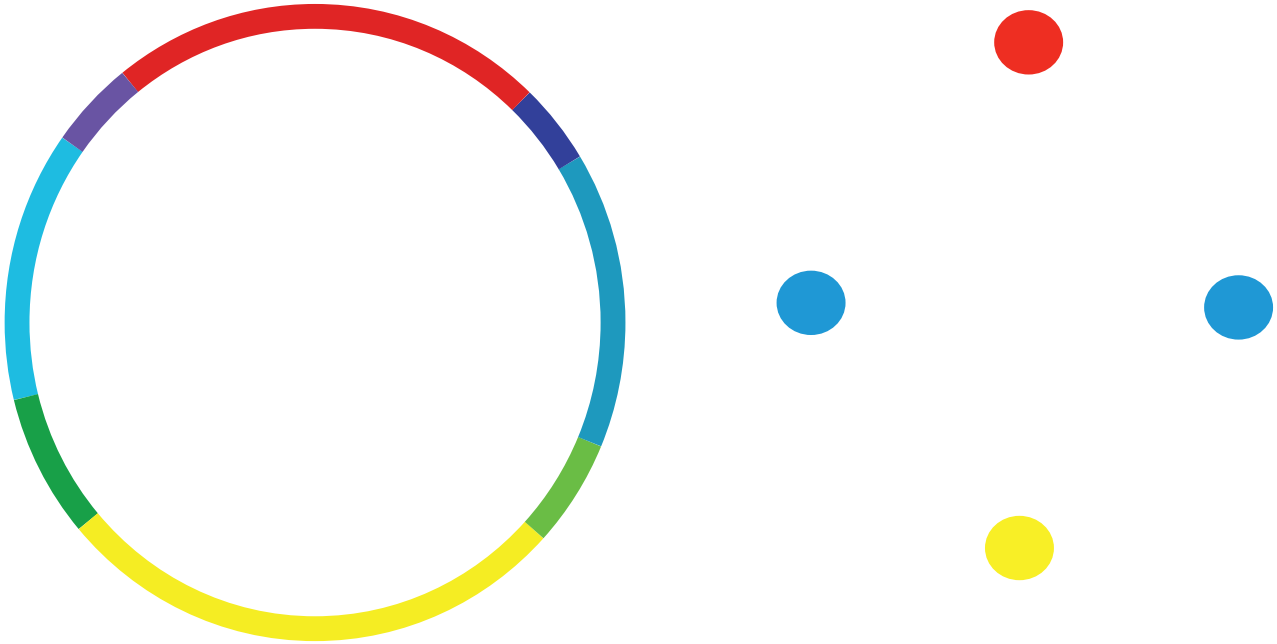


Figure: Create nodes

Nerve complex example

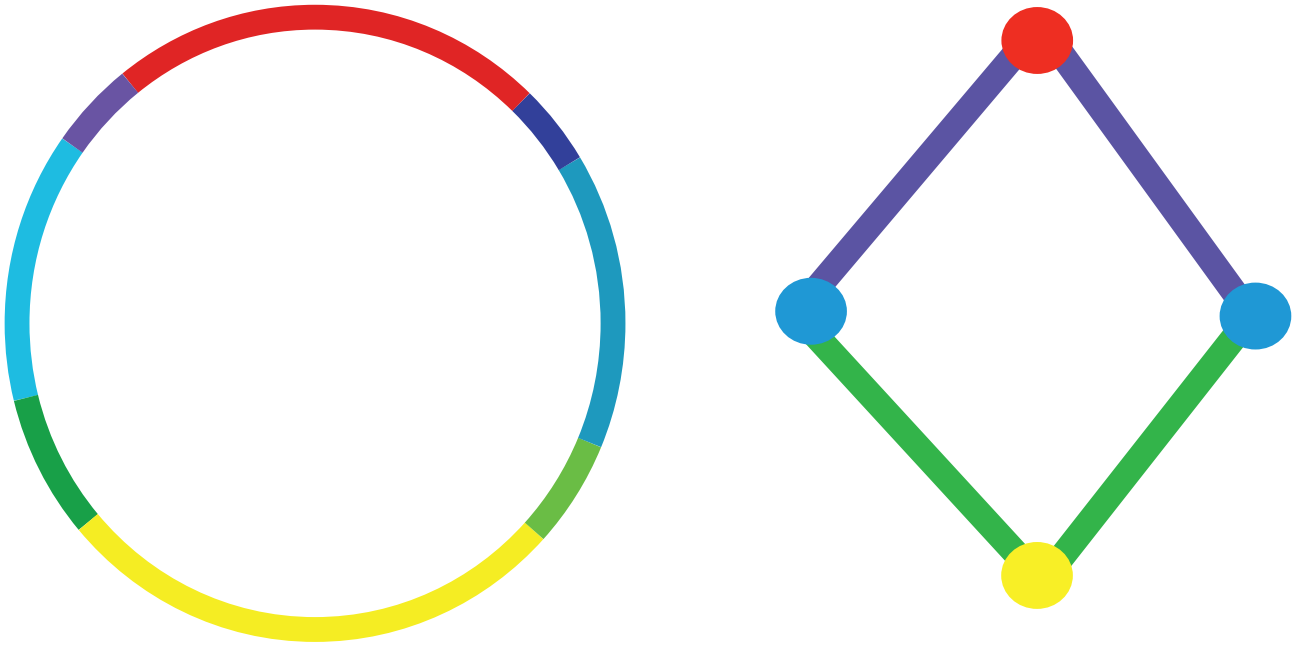


Figure: Create edges, that gives a Nerve complex (graph)

Nerve of Seven Bridges of Königsberg

Figure: Nerve graph of Seven Bridges of Könisberg

Point cloud data

- Now given point cloud data $\mathcal{X} = \{x_1, \ldots, x_n\}$, and a covering $V = \{U_\alpha\}$, where each U_α is a cluster of data
- Build a simplicial complex (Nerve) in the same way, but components replaced by clusters

Mapping

- How to choose coverings?
- Create a reference map (or filter) $h : \mathcal{X} \to \mathcal{Z}$, where $\mathcal{Z}$ is a topological space often with interesting metrics (e.g. $\mathbb{R}$, $\mathbb{R}^2$, S^1 etc.), and a covering $\mathcal{U}$ of $\mathcal{Z}$, then construct the covering of $\mathcal{X}$ using inverse map $\{h^{-1}U_\alpha\}$.

Example: 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 $h^{-1}(t)$

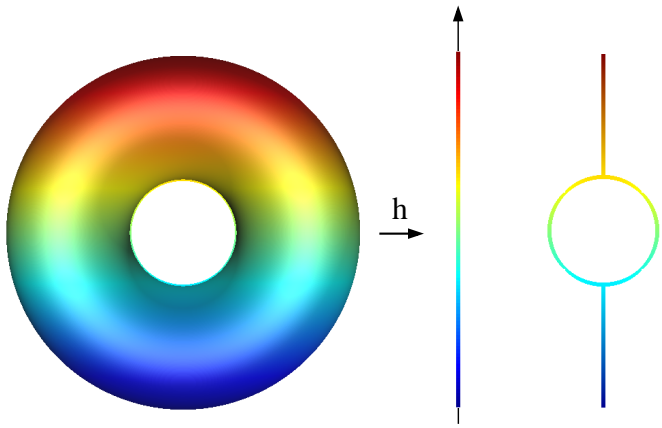


Figure: Construction of Reeb graph; h maps each point on torus to its height.

Mapper: from Continuous to Discrete...

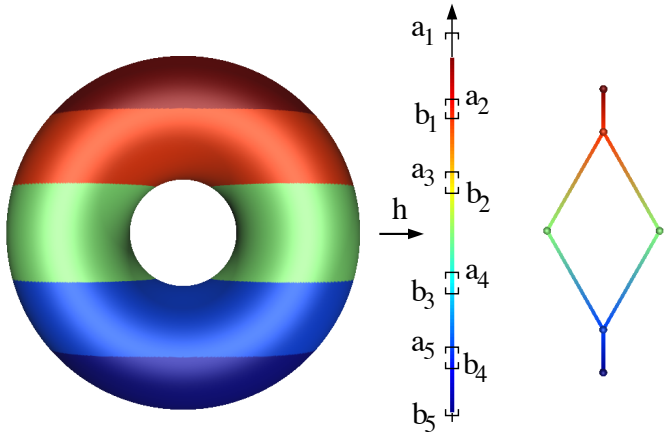


Figure: An illustration of Mapper.

Note:

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

Mapper algorithm

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

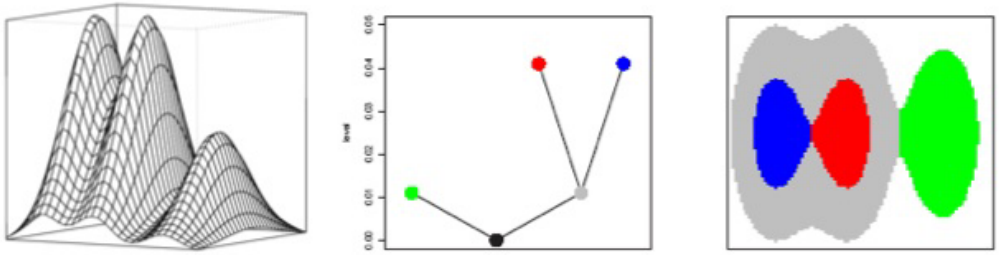
- choose a filter map $h : \mathcal{X} \to \mathcal{Z}$, where $\mathcal{Z}$ is a topological space such as $\mathbb{R}$, S^1, $\mathbb{R}^d$, etc.
- choose a cover $\mathcal{Z} \subseteq \cup_\alpha U_\alpha$
- cluster/partite level sets $h^{-1}(U_\alpha)$ into $V_{\alpha,\beta}$
- graph representation: 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 Reeb Graph from $\mathbb{R}$ to general topological space $\mathcal{Z}$; may lead to a particular implementation of Nerve theorem through filter map h.

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)

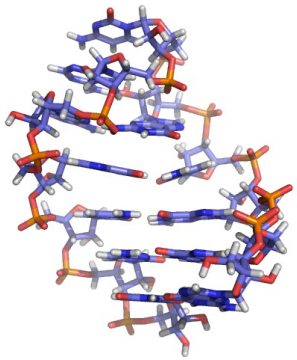


Reference Mapping

Typical one dimensional filters/mappings:

- Density estimators
- Measures of data (ec-)centrality: e.g. $\sum_{x' \in \mathcal{X}} d(x, x')^p$
- Geometric embeddings: PCA/MDS, Manifold learning, Diffusion Maps etc.
- Response variable in statistics: progression stage of disease etc.

Example: RNA Tetraloop



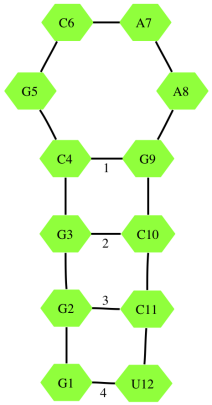


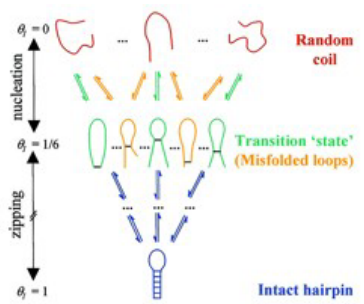
Figure: RNA GCAA-Tetraloop

Biological relevance:

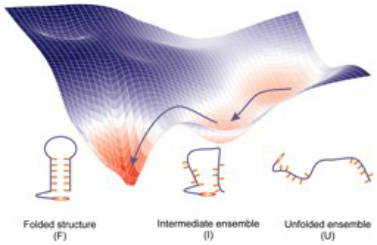
- serve as nucleation site for RNA folding
- form sequence specific tertiary interactions
- protein recognition sites
- certain Tetraloops can pause RNA transcription

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

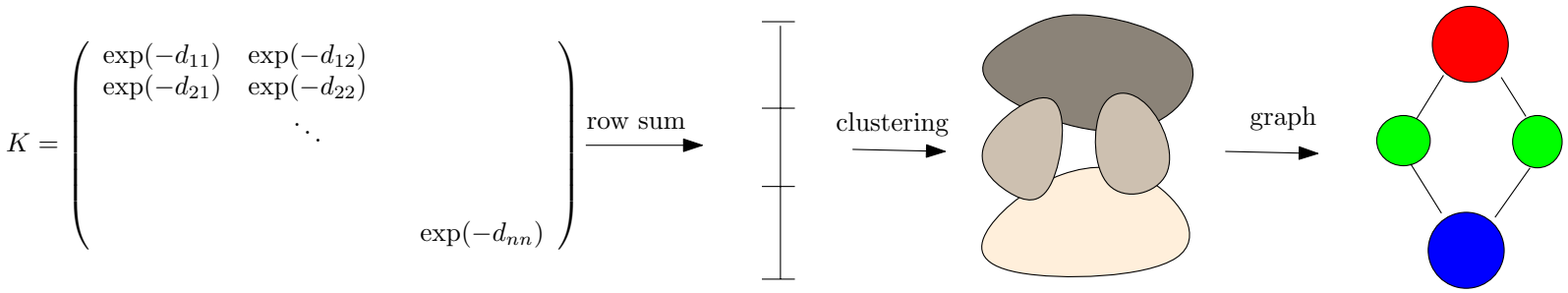


Figure: Mapper Flow Chart

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

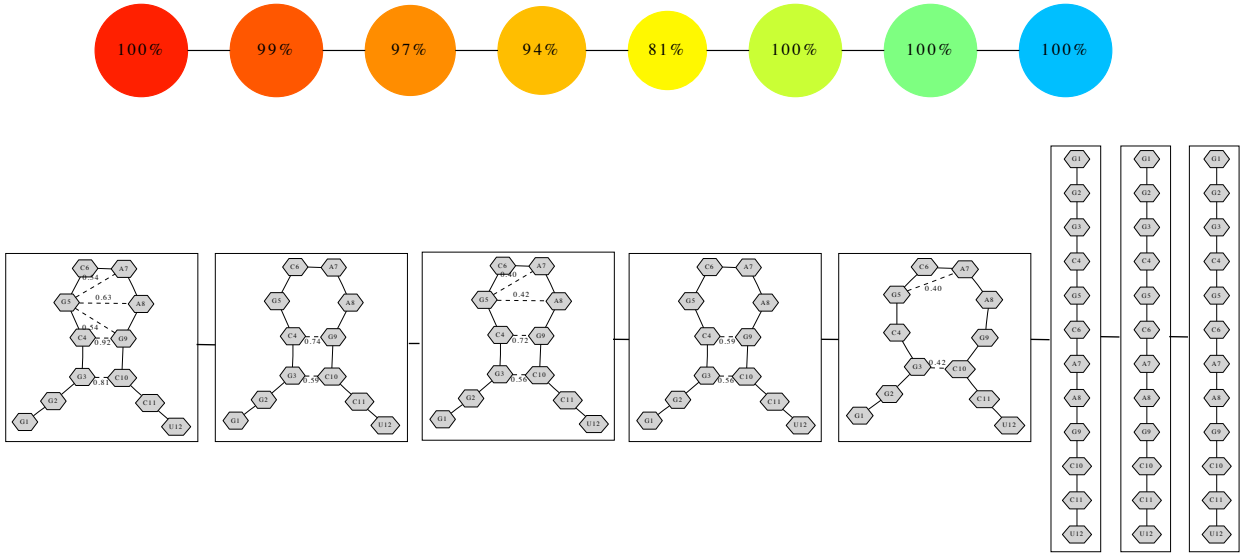


Figure: Unfolding pathway

Mapper output for Refolding Pathways

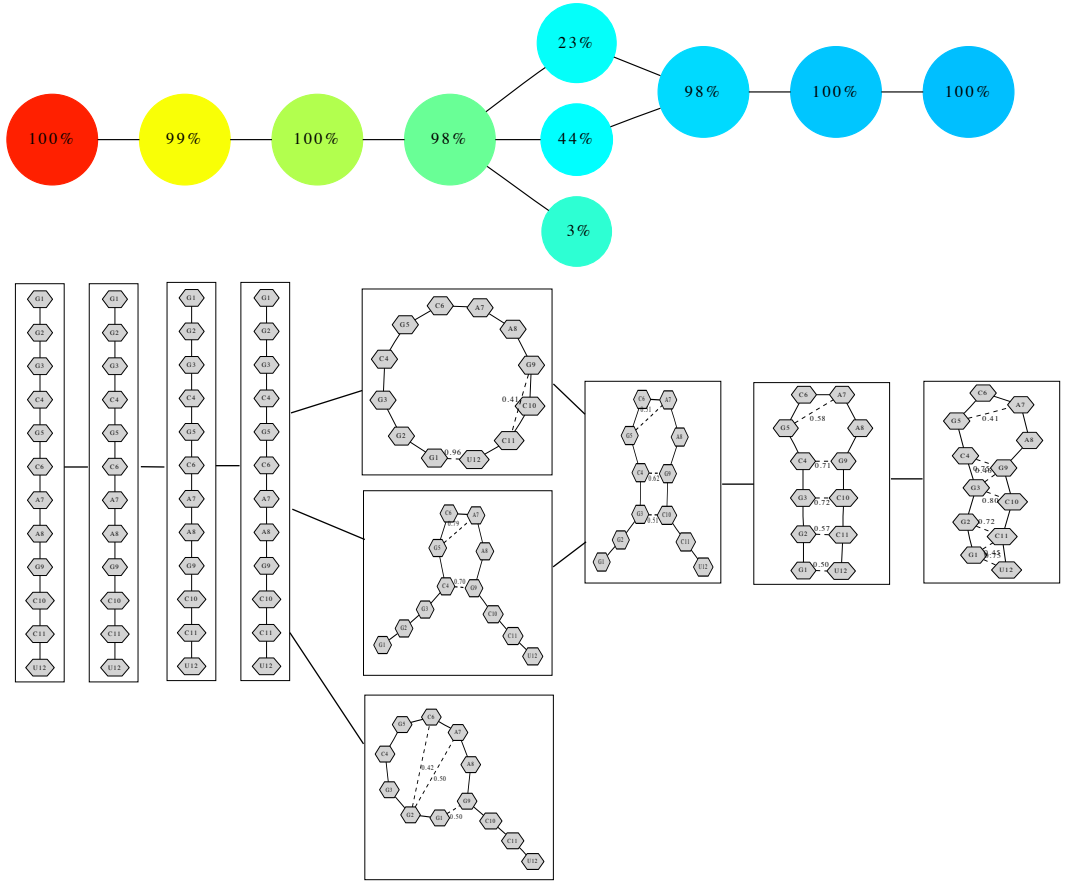


Figure: Refolding pathway

Progression of Breast Cancer: l_2-eccentrality

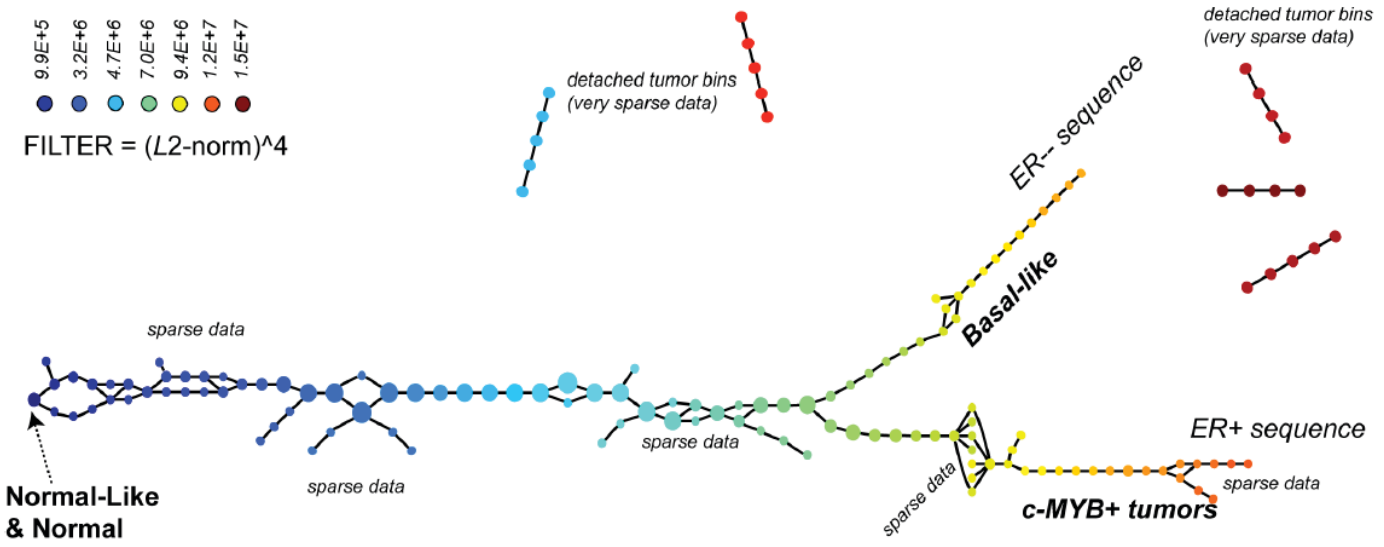


Figure: Monica Nicolau, A. Levine, and Gunnar Carlsson, *PNAS'10*

Cell Cycles

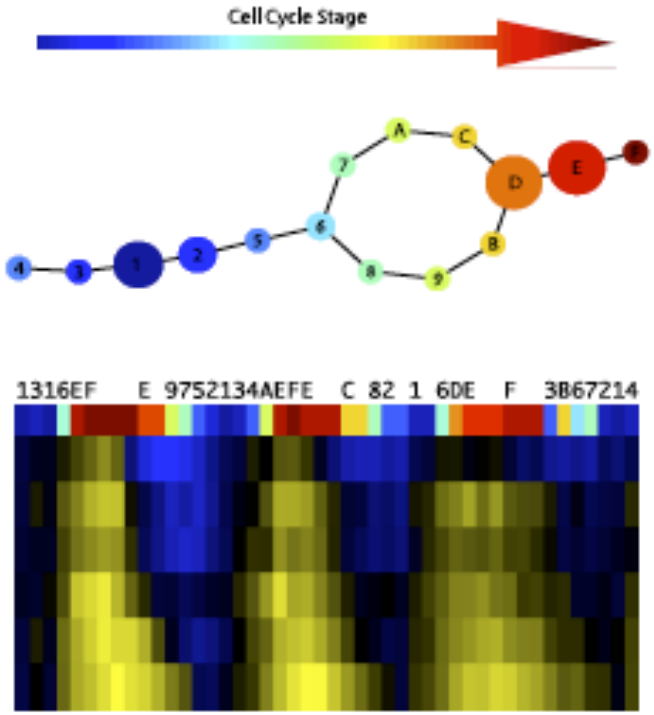


Figure: Cell Cycle Microarray Data, courtesy of M. Nicolau, Nagarajan, G. Singh, Carlsson

Relationships between diabetic, pre-diabetic, and healthy populations

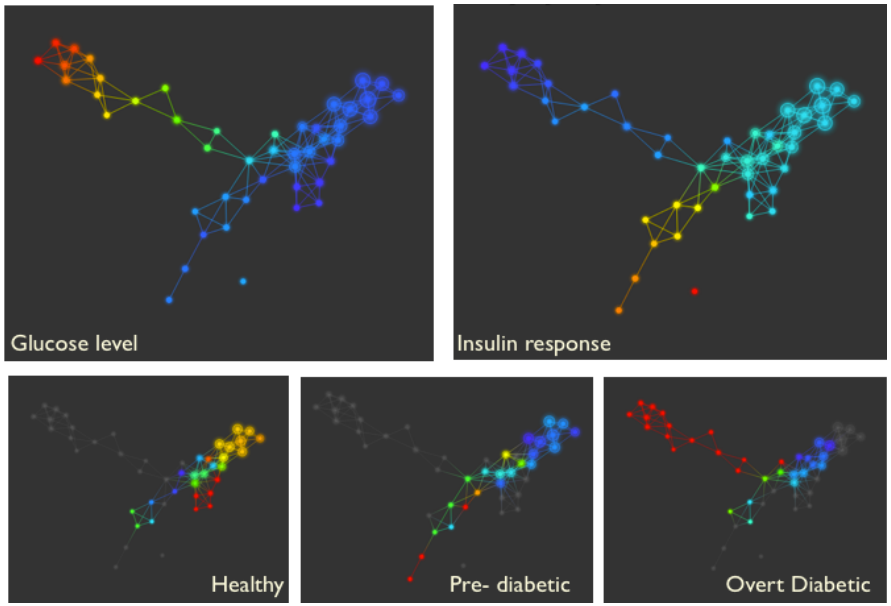
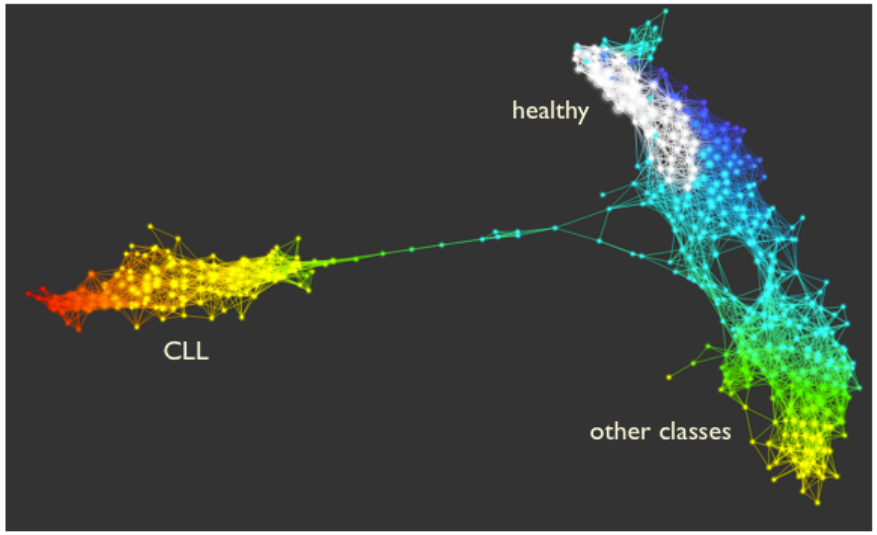


Figure: Miller-Reaven Diabetes Dataset, courtesy of Gunnar Carlsson

Leukemia with gene expression profiles



Data: Gene expression profiles of bone marrow of leukemia patients
Source: PMID 8573112
Columns: 1500 genes
Rows: 1905 patients

Figure: Topological structure of Leukemia: courtesy of Gunnar Carlsson