Topological Methods for Exploring Low-density States on Biomolecular Folding Pathways

Yuan Yao

Stanford University

June 26, 2008
Acknowledgements

Collaborators:

- Biology: Xuhui Huang, Greg Bowman, Vijay Pande
- Computer Science: Jian Sun, Leo Guibas
- Mathematics: Michael Lesnick, Gurjeet Singh, Gunnar Carlsson

Thanks to

- Michael Levitt
- Wing-Hung Wong
- Nancy Zhang
A motivating example: RNA 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

- experiments: no structural information
- computer simulations at full-atom resolution:
  - existence of intermediate states
  - if yes, what’s the structure?
SREMD Simulations

[Bowman, Huang, Y., Sun, ... Vijay. JACS, 2008, to appear]

- 2800 SREMD (Serial Replica Exchange Molecular Dynamics) simulations with RNA hairpin (5’-GGGCAGCAAGCCU-3’)
- 389 RNA atoms, ∼4000 water and 11 Na⁺
- SREMD random walks in temperature space (56 ladders from 285K to 646K) with molecular dynamic trajectories
- 210,000 ns simulations with ∼105,000,000 configurations
- Unfortunately, sampling still not converged!
Challenges for Data Analysis

- Massive data: $\sim 100M$ samples
- High dimensionality: $12K$ Cartesian coordinates
- Looking for a needle in a haystack:
  - intermediates/transition states of interests are of low-density
  - folded/unfolded states are dominant
- Samples are not in equilibrium distribution
Dimensionality reduction: Contact maps

- 12 residues for each conformation
- two nonadjacent residues are in **contact** if their nearest atoms are within 3 Å
- every configuration as a **undirected graph**, described by 55-bit string

![Diagram](image)

**Figure:** (a) NMR structure of the GCAA tetraloop. (b) Contact map for the native state. Bases are numbered from 1 to 12 and native basepair contacts are numbered 1-4.
Further discussions on contact maps

- Contact maps faithfully represent the spatial relations between stem base-pairs.
- Stem base-pair formation is crucial to characterize the structures of intermediate states.
- Other representation like RMSD is too noisy due to the heterogeneity in loop shapes.
- Distance metric between contact maps: Hamming distance.
- Such a metric is too coarse for nonlinear dimensionality reduction methods (e.g. ISOMAP \cite{Das, et al. PNAS, 2006, 103:9885-9890}) to find reaction coordinates.
Needle through magnifying glasses: Conditional density functions

Conditioning on the region where intermediate states may host:
- folding/unfolding events
- biased toward the target states (folded/extended)

Note: applicable to non-equilibrium distributed data.

- 760 unfolding events;
- 550 folding events;
Our strategy

**Problem:** How to separate sparse intermediates from dense uninterested structures?

**Solution**

- **stratify data into density level sets, and**
- **cluster on each level set**

**But,** can we organize those clusters in a systematic way?
- Yes, **Morse theory** in mathematics provides an inspiration...
Morse Theory and Reeb graph

- a nice (Morse) function: \( h : \mathcal{X} \rightarrow \mathbb{R} \), 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, or Hartigan trees
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 $T \subseteq \bigcup_{\alpha} U_{\alpha}$
- 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
Mapper with density filters in biomolecular folding

In biomolecular folding

- **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

$$K = \begin{pmatrix}
\exp(-d_{11}) & \exp(-d_{12}) \\
\exp(-d_{21}) & \exp(-d_{22}) \\
\vdots & \\
\exp(-d_{nn})
\end{pmatrix}$$

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
Figure: Unfolding pathway
Mapper output for Refolding Pathways

Figure: Refolding pathway
The two intermediate states, are on-pathways; the inner base-pair formation is easier in proceeding than backing (.15/.07), while the end base-pair formed more reluctant (.12/.09)

Note that this is not a Markov State Model.
Biological Suggestions from Mapper Results


- Folding and unfolding follows different pathways
- For folding pathways, there are multiple intermediate states
  - a dominant one with inner/closing stem base-pair formed
  - a less dominant one with outer/end stem base-pair formed
- This in the first time provides structural evidence in support of multistate hypothesis on folding pathways
Open problems and future directions

- Only static information is used, how to incorporate kinetic information?
- Combine geometric embedding with topological methods for better characterization of reaction coordinates?
- Toward a new generation of transition networks (Markov State Models)?
  - Mapper may characterize both metastable states and intermediate/transition states on different density/energy level sets
  - Traditional transition networks are based on metastates, which can be inferred from Mapper results