UNIVERSITY of DELAWARE

Enabling Scalable Data Analysis for Large Computational Structural Biology Datasets on Distributed Memory Systems

Michela Taufer

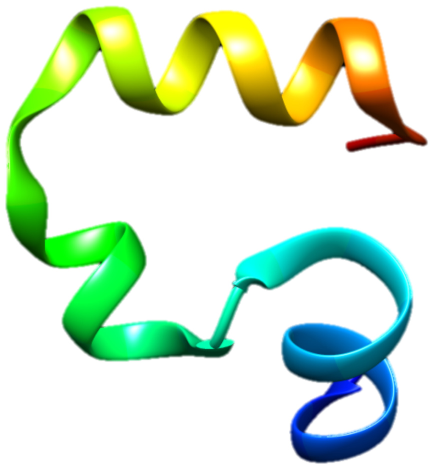
Global Computing Laboratory

Computer and Information Sciences

University of Delaware

In-situ Analysis

- The perfect in-situ analysis algorithm:
  - Avoids the need for moving data
  - Uses a limited amount of memory
  - Executes sufficiently fast

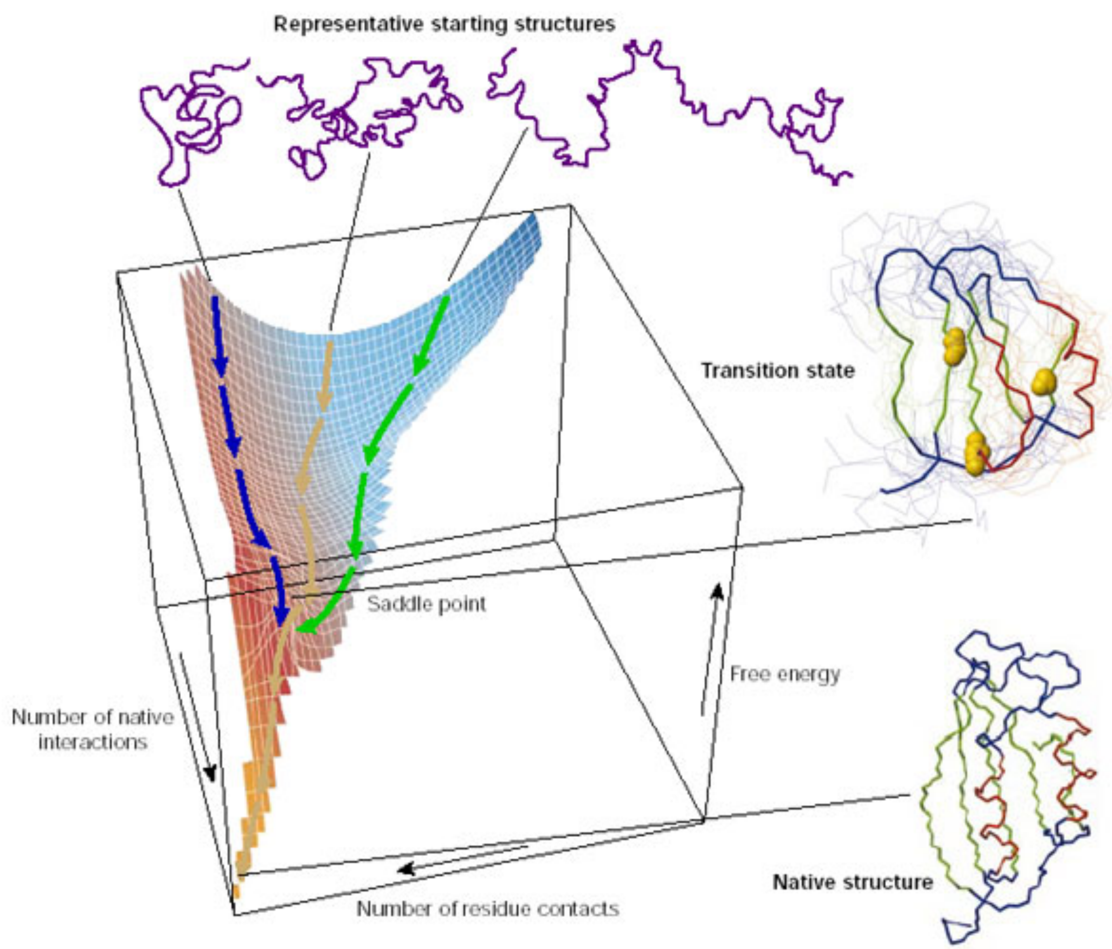


Dataset NleNle - PDB ID 2F4K
from Vijay Pande group

Can we perform in-situ analysis on trajectories generated in protein folding, prediction, or refinement simulations?

Protein Folding Process

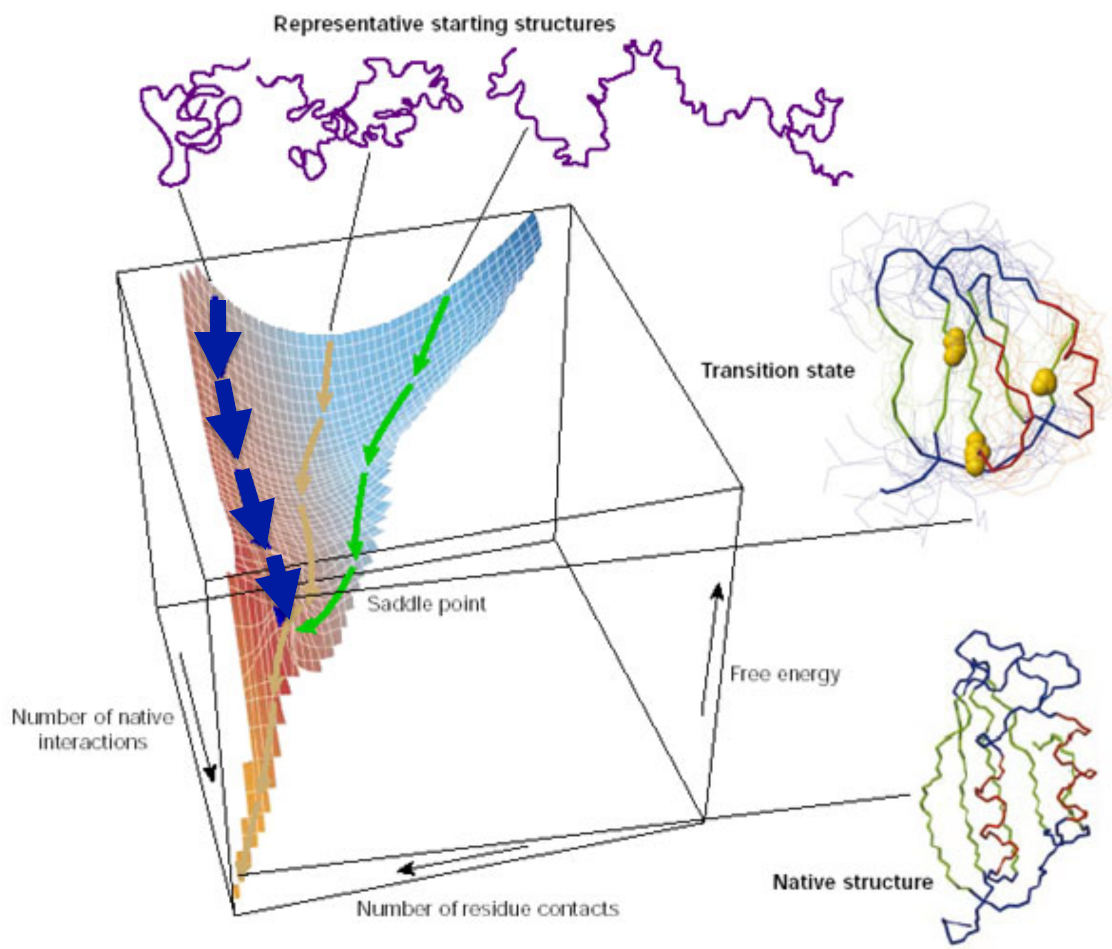
- Start from unfolded conformations of a protein with correct chemical bonds but random torsion angles
- Search for a conformation close to the observed native (folded) conformation



Reynaud, E. (2010) Protein Misfolding and Degenerative Diseases. Nature Education 3(9):28

Protein Folding Process

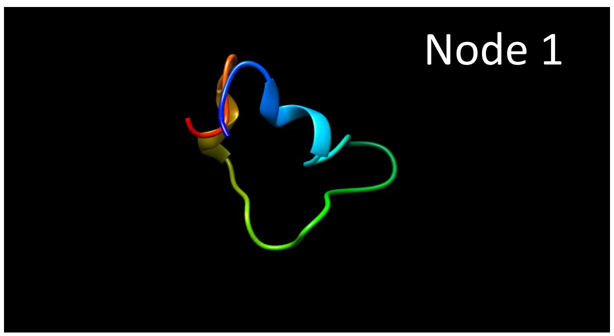
- Start from unfolded conformations of a protein with correct chemical bonds but random torsion angles
- Search for a conformation close to the observed native (folded) conformation

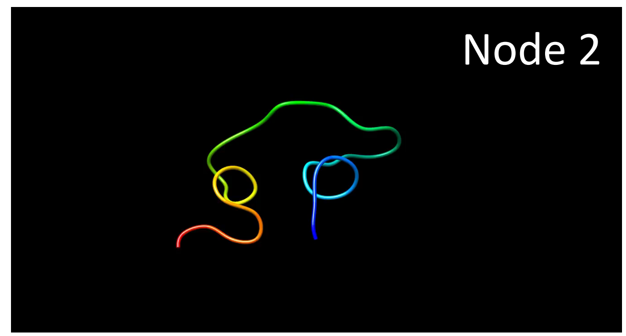


Reynaud, E. (2010) Protein Misfolding and Degenerative Diseases. Nature Education 3(9):28

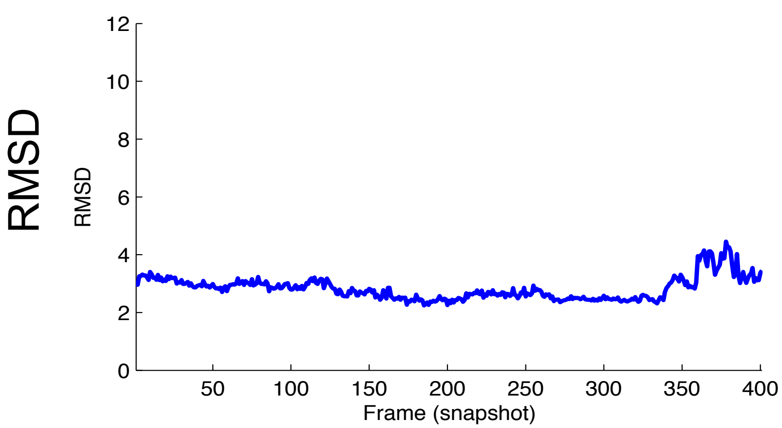
Scientific Problem

- Cluster folding trajectories into recurrent patterns based on geometrical variations in time of the folding protein frames

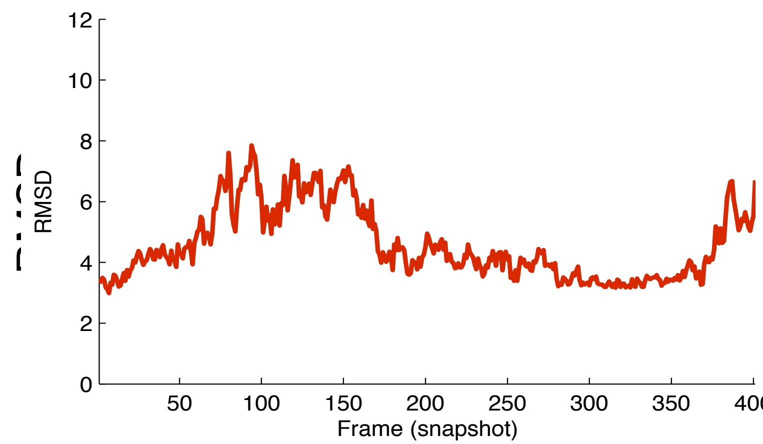




Trajectory segment 1

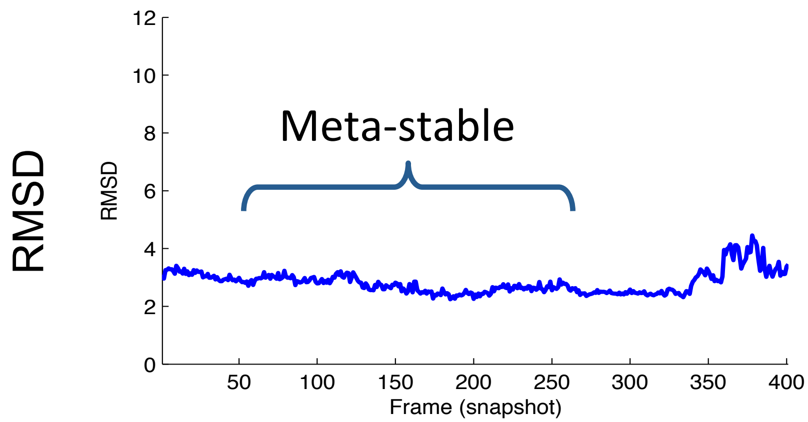


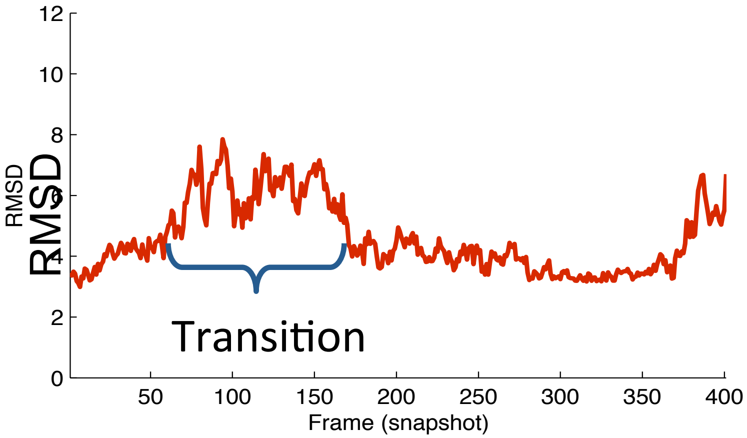
Trajectory segment 2



Scientific Problem

- Cluster folding trajectories into recurrent patterns based on geometrical variations in time of the folding protein conformations
- Intra-trajectory analysis -> identify meta-stable and transition stages within trajectory



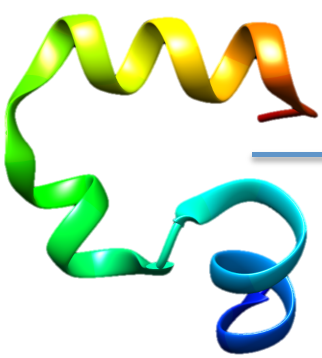


Limits of Current Practice

- Centralized clustering analysis [Phillips et al. 2013]
  - Wait for a job to end before to move its data (trajectory segment) to a centralize server
  - Does not scale when the dataset is large
  - May end up wasting computing resources e.g., by trying to further fold proteins that are already in a stable condition

From Protein Conformations to Metadata

- From a single protein conformation to a 3D point capturing the atom distances within the frame



Conformation
(Data from F@H)

distance matrix (DM)

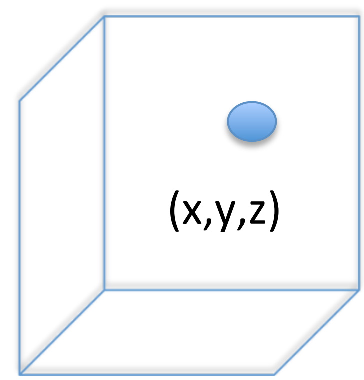
atoms / atoms

	1	2	3	4	...	N
1	0.0	2.0	1.8	3.2	...	10.4
2	2.0	0.0	2.3	2.1	...	9.3
3	1.8	2.3	0.0	0.7	...	9.4
4	3.2	2.1	0.7	0.0	...	8.9
...	...	...	...	...	0.0	...
N	10.4	9.3	9.4	8.9	...	0.0

	1 (x)	2 (y)	3 (z)
1	1.0	2.3	1.4
2	2.0	0.4	1.3
3	1.8	2.3	2.0
4	1.2	3.1	1.7
...	...	...	...
N	4.4	2.3	3.4

multi-dimensional scaling

three leading eigenvalues

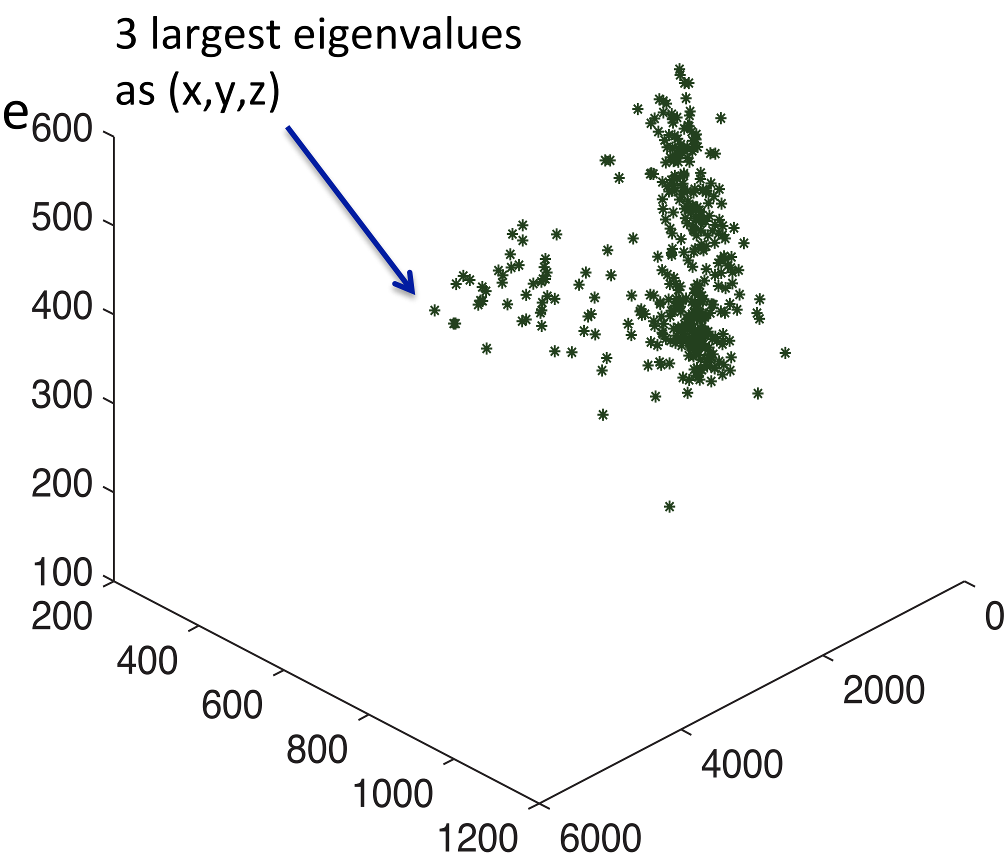


metadata

B. Zhang, T. Estrada, P. Cicotti, and **M. Taufer**. Enabling In-situ Data Analysis for Large Protein-folding Trajectory Datasets. In *IPDPS*, 2014

From Metadata to Scientific Knowledge

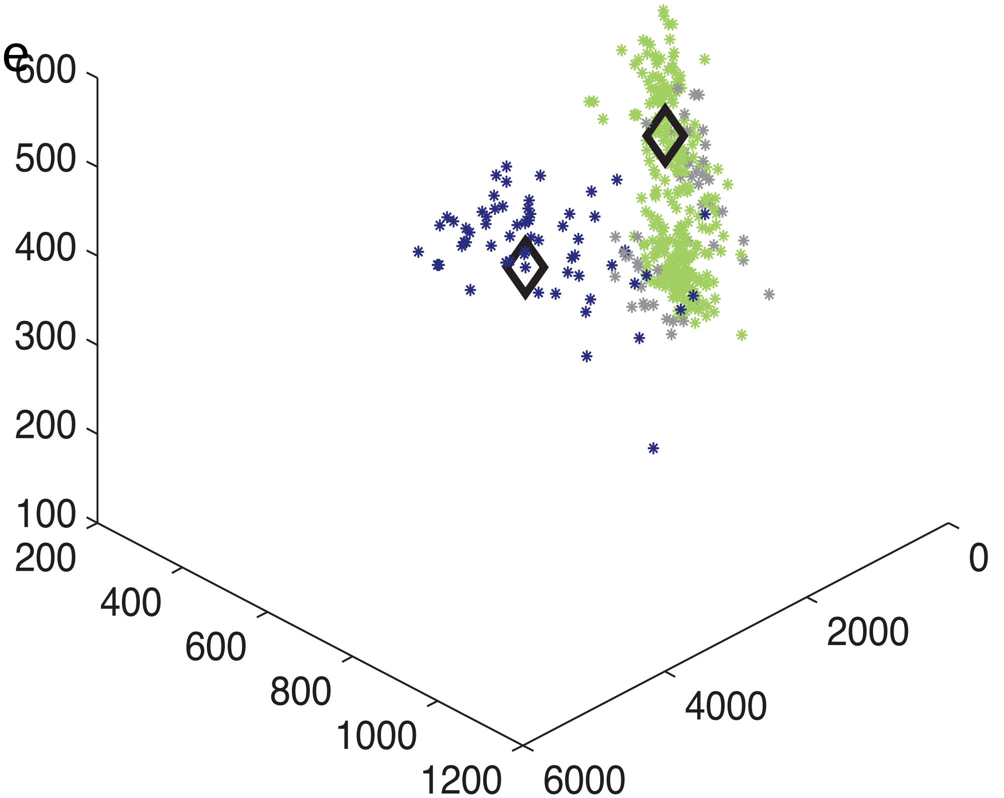
Given a set of confirmations and their 3D points in a segment of the trajectory:



From Metadata to Scientific Knowledge

Given a set of confirmations and their 3D points in a segment of the trajectory

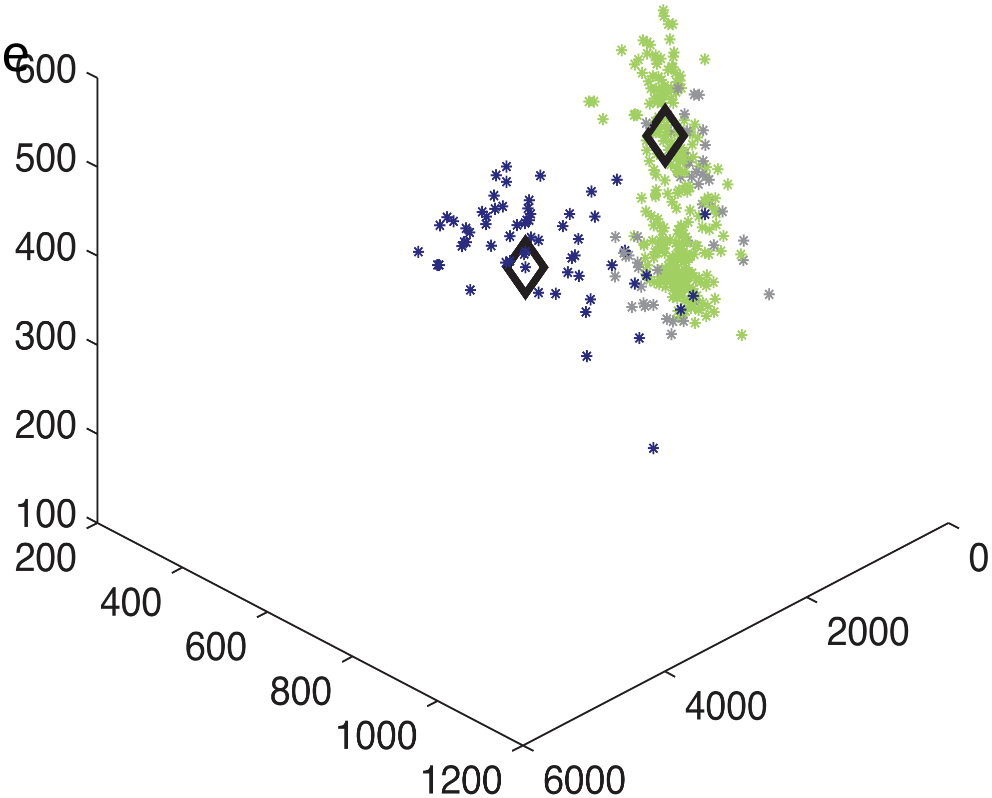
- Partition the 3D points into 2 clusters using fuzzy c-means (with c = 2)



From Metadata to Scientific Knowledge

Given a set of confirmations and their 3D points in a segment of the trajectory:

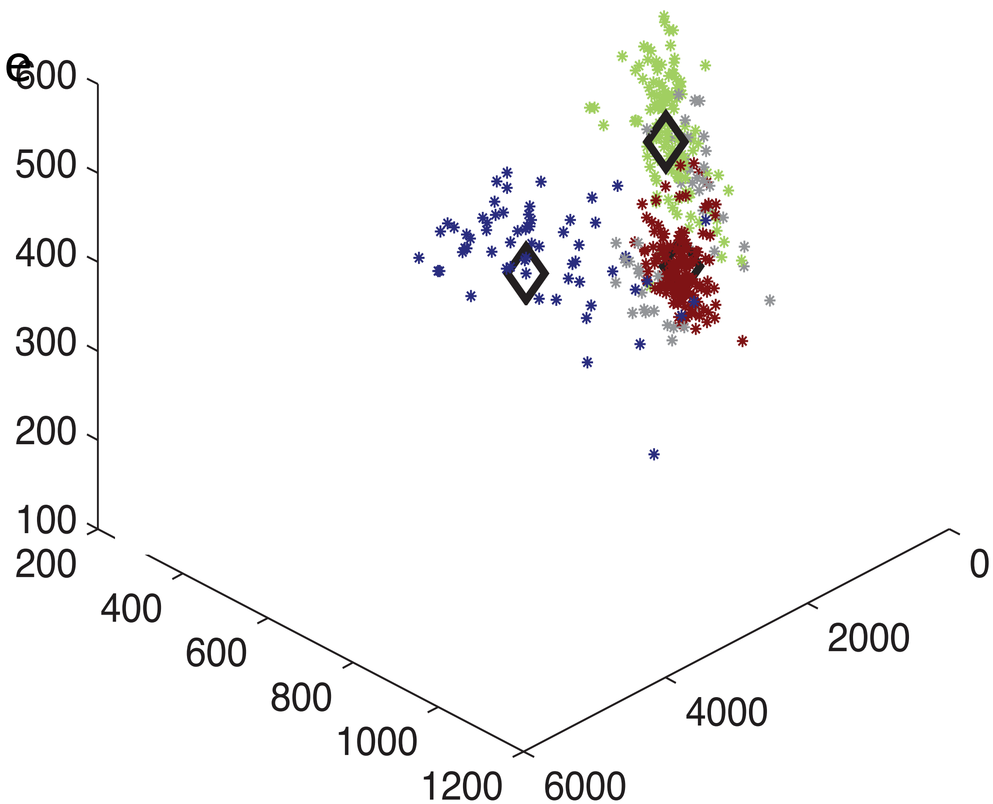
- Partition the 3D points into 2 clusters using fuzzy c-means (with c = 2)
- Test the equality of the 2 clusters using Welch's t-test with p-value < 0.01



From Metadata to Scientific Knowledge

Given a set of confirmations and their 3D points in a segment of the trajectory:

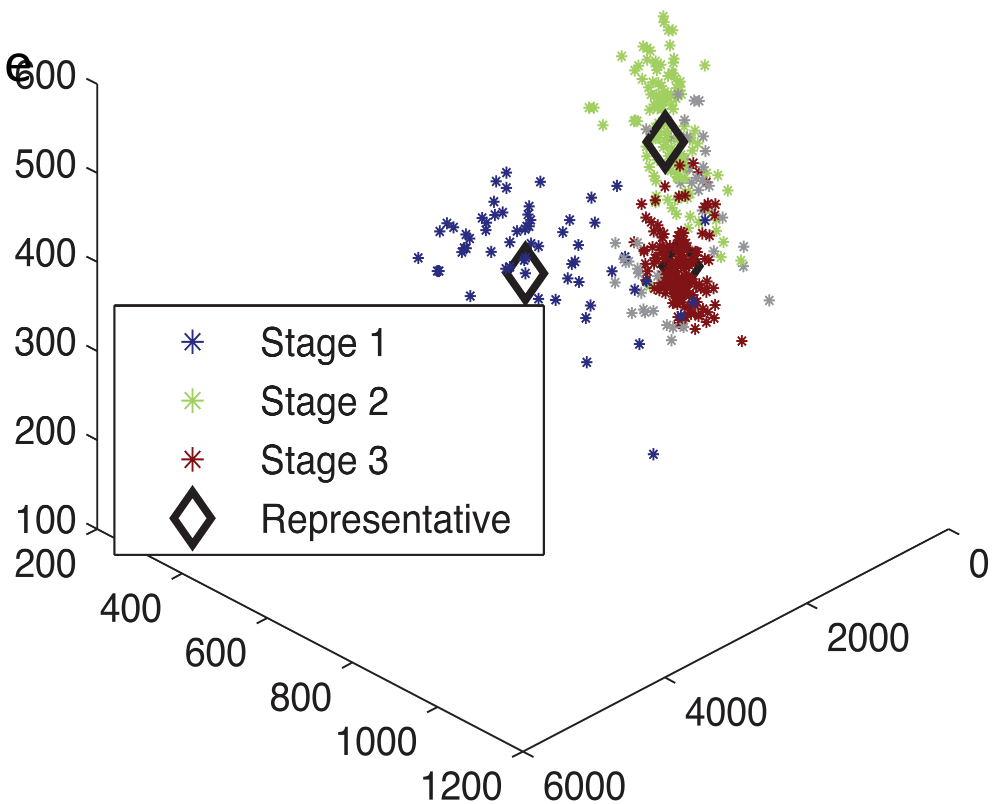
- Partition the 3D points into 2 clusters using fuzzy c-means (with c = 2)
- Test the equality of the 2 clusters using Welch's t-test with p-value < 0.01
- Iterative partition on the cluster with more points



From Metadata to Scientific Knowledge

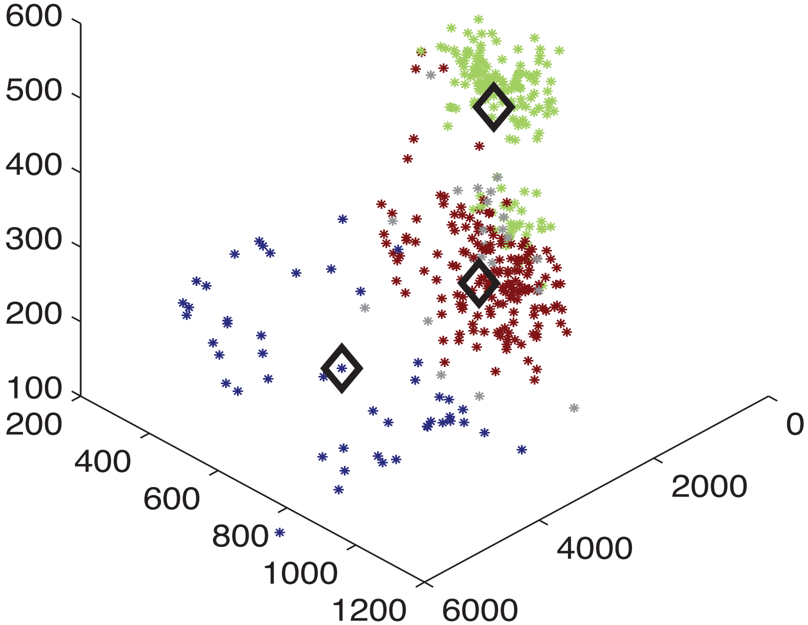
Given a set of confirmations and their 3D points in a segment of the trajectory:

- Partition the 3D points into 2 clusters using fuzzy c-means (with c = 2)
- Test the equality of the 2 clusters using Welch's t-test with p-value < 0.01
- Iterative partition on the cluster with more points
- Finish when the 2 clusters are equal



From Metadata to Scientific Knowledge

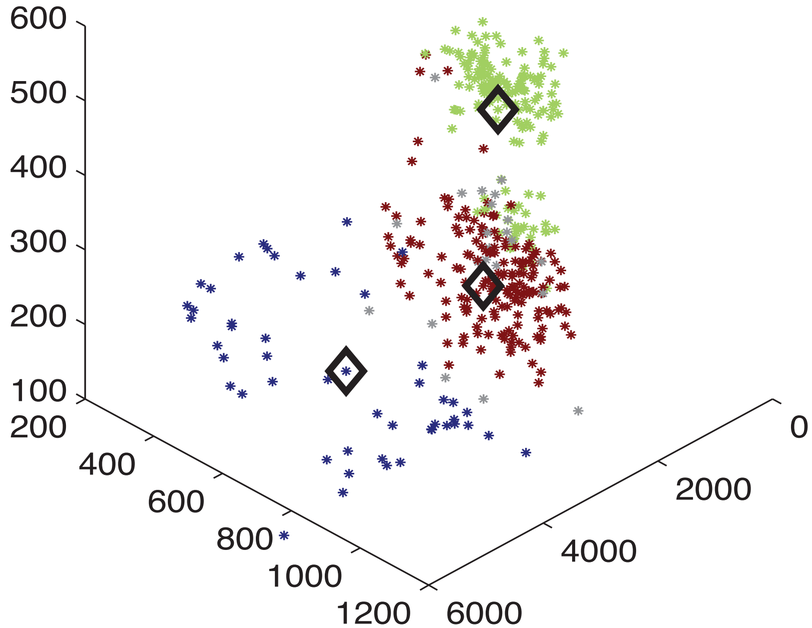
- Intra-trajectory analysis of an ensemble of 400-conformations containing one meta-stable stage followed by a transition stage and another meta-stable stage



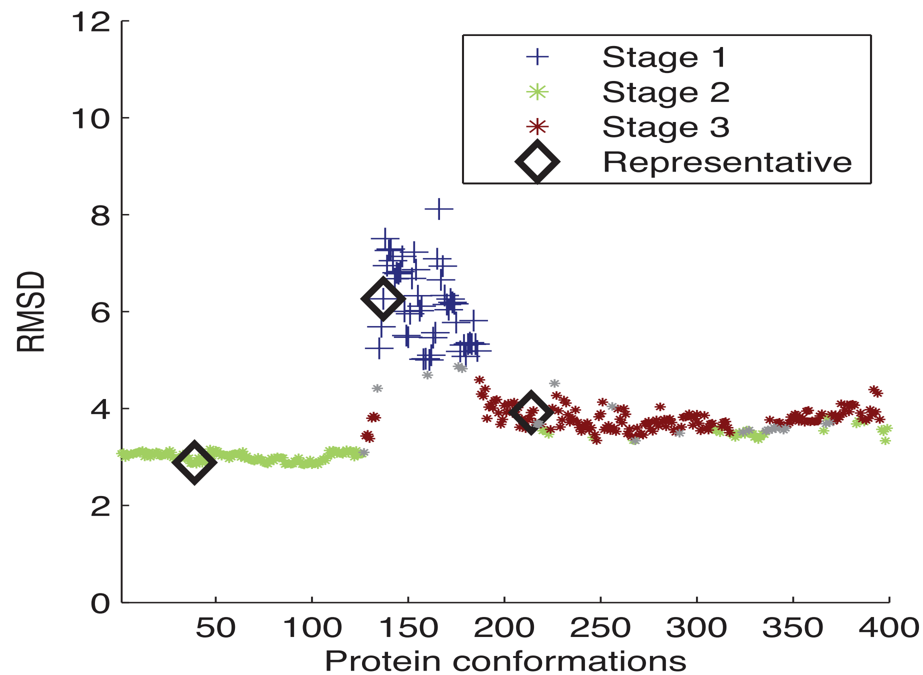
Our clustering identifies two stable stages and one transition stage

From Metadata to Scientific Knowledge

- Intra-trajectory analysis of an ensemble of 400 conformations containing one meta-stable stage followed by a transition stage and another meta-stable stage



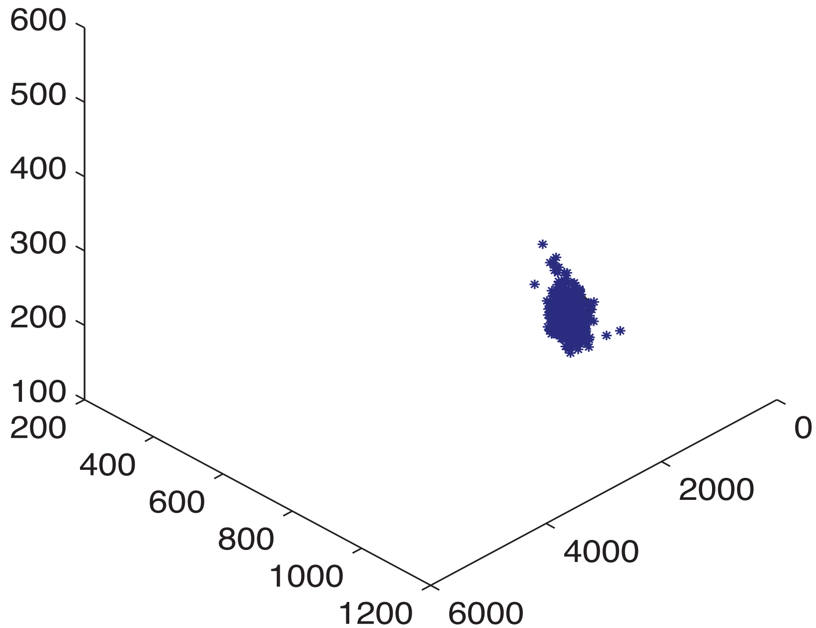
Our clustering identifies two stable stages and one transition stage



RMSDs of protein conformations identify the same three stages

From Metadata to Scientific Knowledge

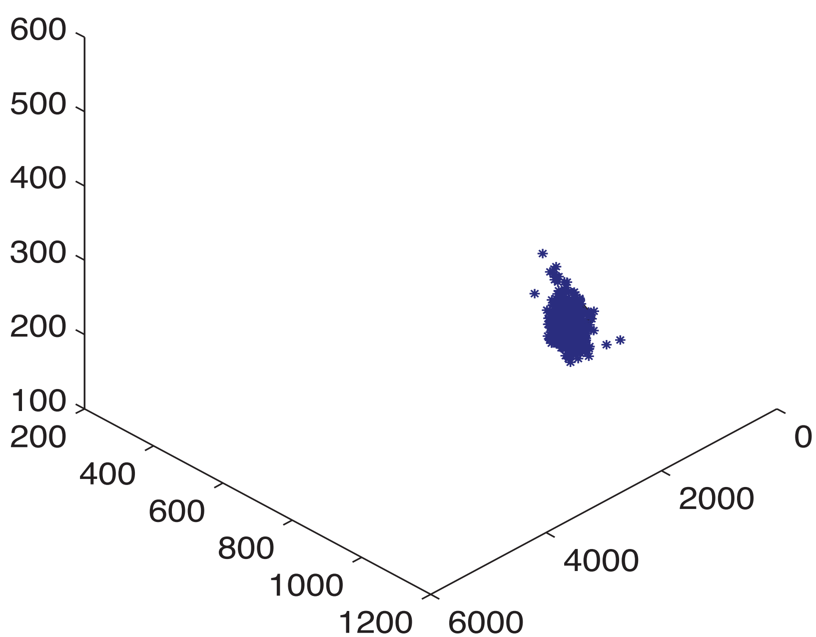
- Intra-trajectory analysis of an ensemble of 400-conformations containing one meta-stable stage only



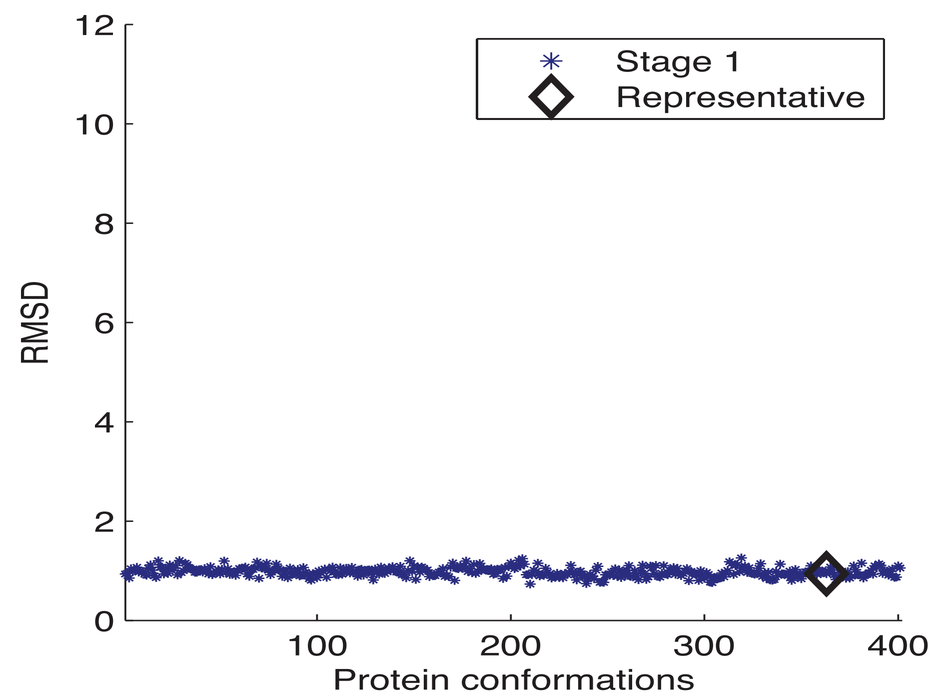
Our clustering identifies one single stage

From Metadata to Scientific Knowledge

- Intra-trajectory analysis of an ensemble of 400-conformations containing one meta-stable stage only



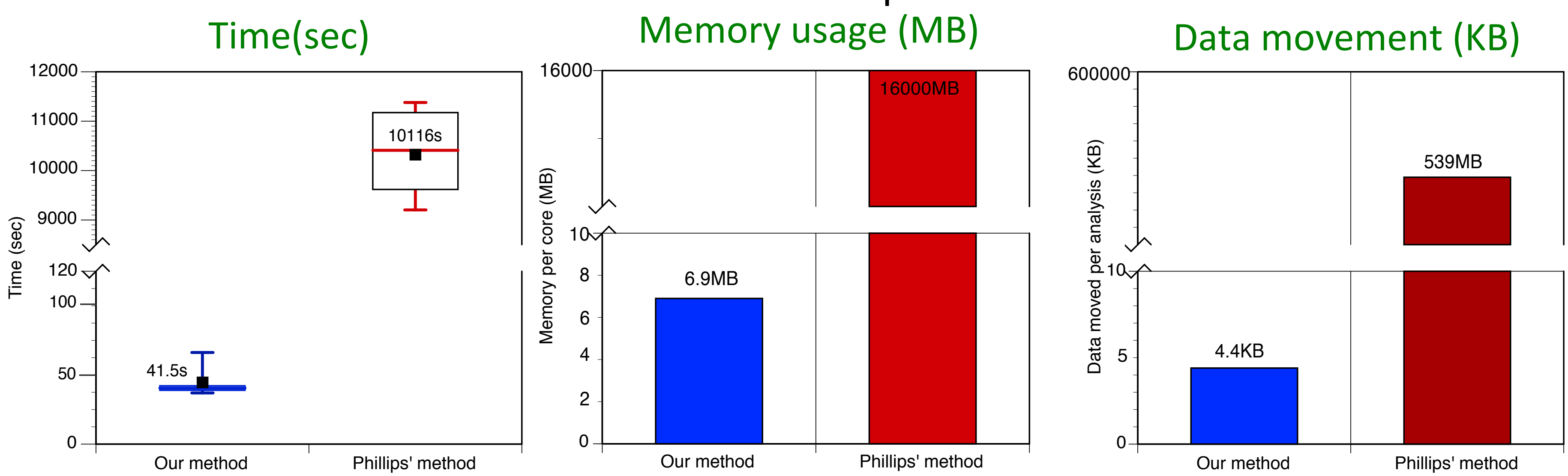
Our clustering identifies one single stage



RMSDs of protein conformations identify one single stage

Performance

- Compare our approach with traditional clustering method proposed by Phillips et al.
  - Folding trajectories of villin headpiece subdomain (HP-35 NleNle)
  - Parallel MATLAB on 256 Gordon compute cores



Our method performs orders of magnitude better in terms of time, memory usage and data movement

Lessons Learned and Future Directions

- The distributed analysis of structural biology data is feasible and scalable
- Our approach transforms data properties into metadata concurrently and extracts scientific insights from metadata with small data movement
- Our approach makes the in-situ analysis feasible by:
  - By avoiding the need for moving data, using a limited amount of memory, and executing sufficiently fast

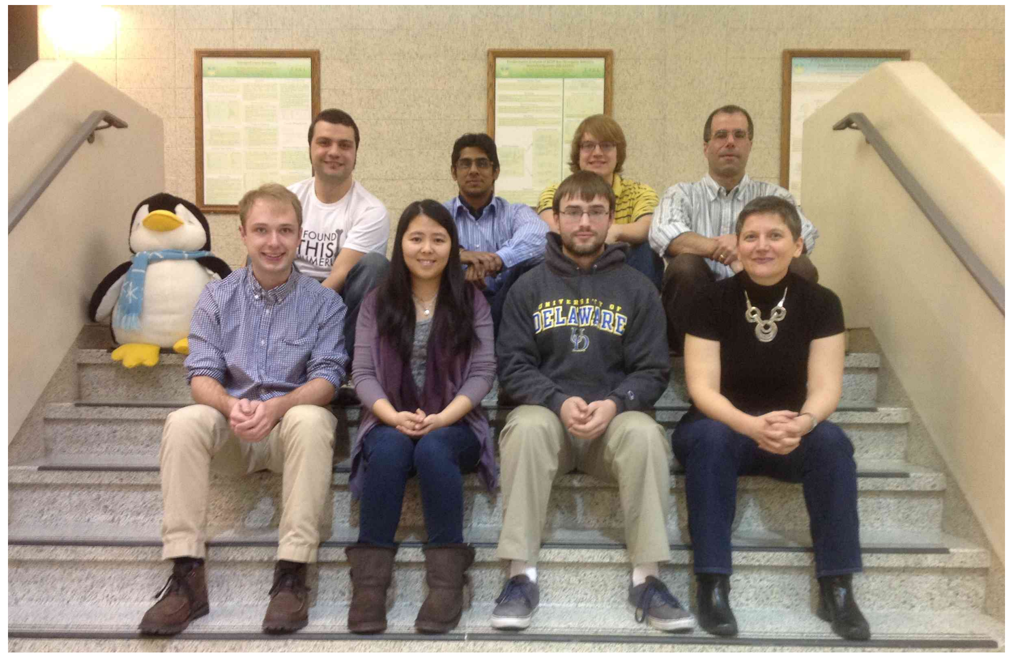
Can we extend our approach to larger scales, i.e., larger datasets and more diverse datasets?

Acknowledgments

Global Computing Lab:
- Boyu Zhang
- Travis Johnston

Collaborators:
- Trilce Estrada (UNM)
- Pietro Cicotti (SDSC)
- Roger Armen (TJU)

Special thanks to:
- D.E. Shaw group
- Vijay Pande group (Stanford)
- Docking@Home volunteers
- Folding@Home volunteer



Sponsors:

XSEDE Extreme Science and Engineering Discovery Environment

NSF

NVIDIA.