In Situ Data Analytics and Indexing of Protein Trajectories

Michela Taufer

Computer and Information Sciences

University of Delaware

Newark, Delaware, USA

Acknowledgements

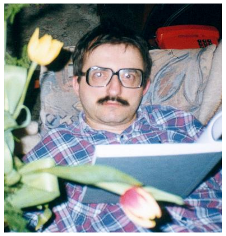
Sponsors:

Collaborators:

Travis J.

Boyu Z.

Trilce E.



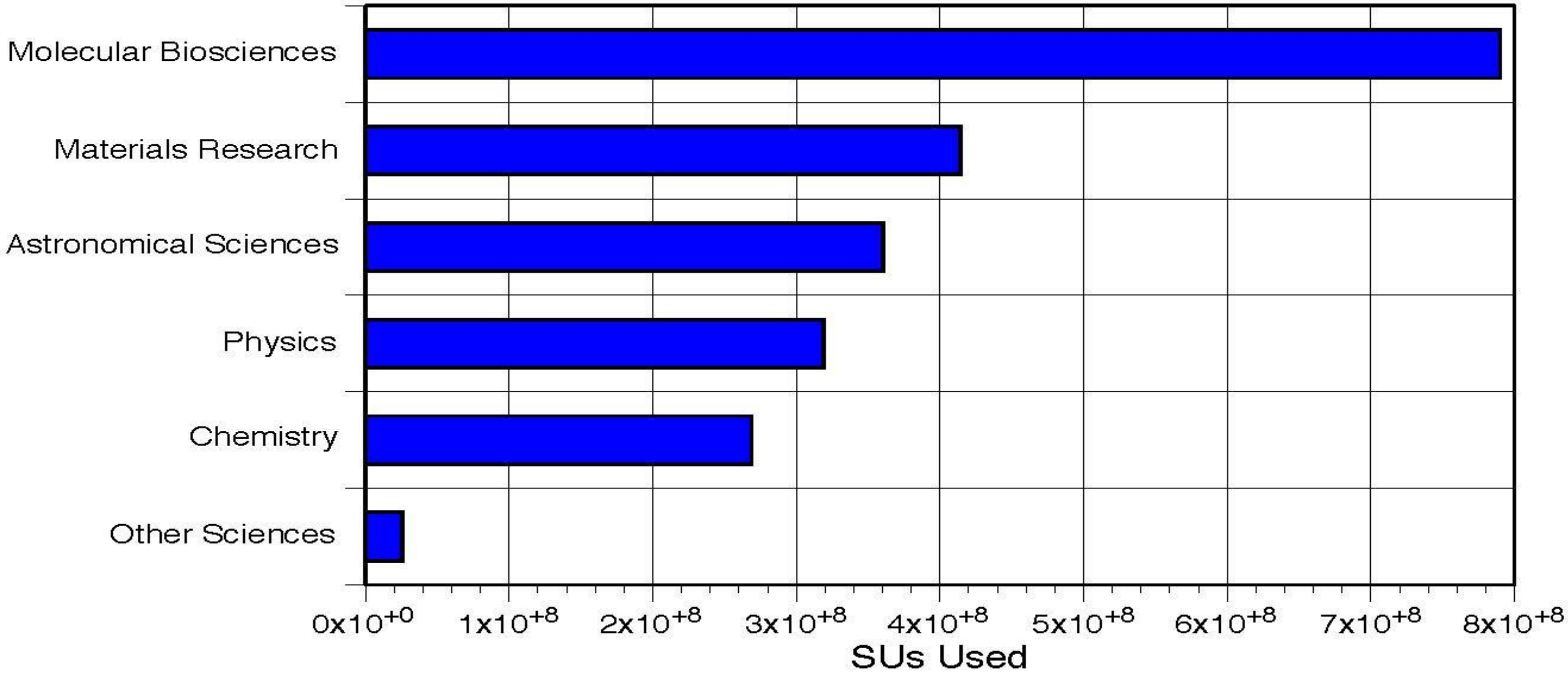
Adam L.

Silvia C.

The GCLab@UD

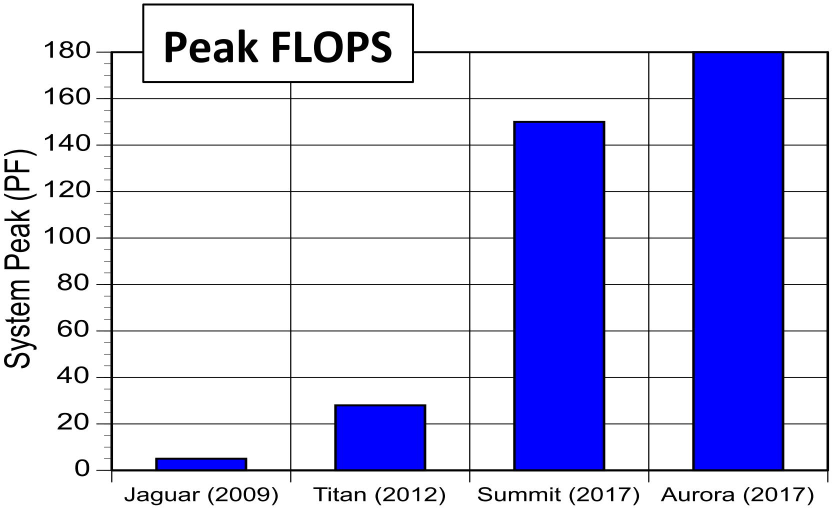
MD simulations are alive and kicking!

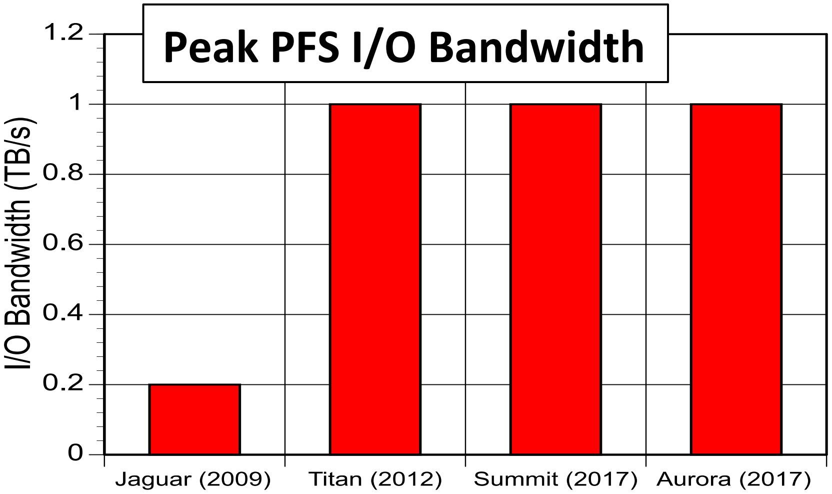
XSEDE SUs used by type of targeted science over the past 6 months (March 1, 2016 - August 31, 2016)



Four of the top 10 XSEDE users run molecular simulations (i.e., Schulten at UIUC, Feig at Michigan State U, Voth at U Chicago, and Case at Rutgers U)

Challenges at the Extreme Scale





MD simulations today:

- Save all the data to analyze later!

MD simulations at exascale:

- Analyze data as they are generated
- Save only what is really needed!
- Data must be easy to retrieve and reuse

We must change how we run our simulaitons at the exascale

Data from Lucy Nowell, DoE Program Director (DoE Workflow Workshop, Rockville, MD, USA. April 20-21, 2015)

Perspectives

The scientist:

“Storage technologies are advancing […] and it is really not clear at all [to me] that especially distributed storage platforms would not be able to handle […] petabyte data sets”

Anonymous Feedback

The computer architect:

“[…] there will be burst buffers on the DOE machines which will give applications much faster I/O […]”

Anonymous Feedback

Burst Buffers

Many have heard about it,

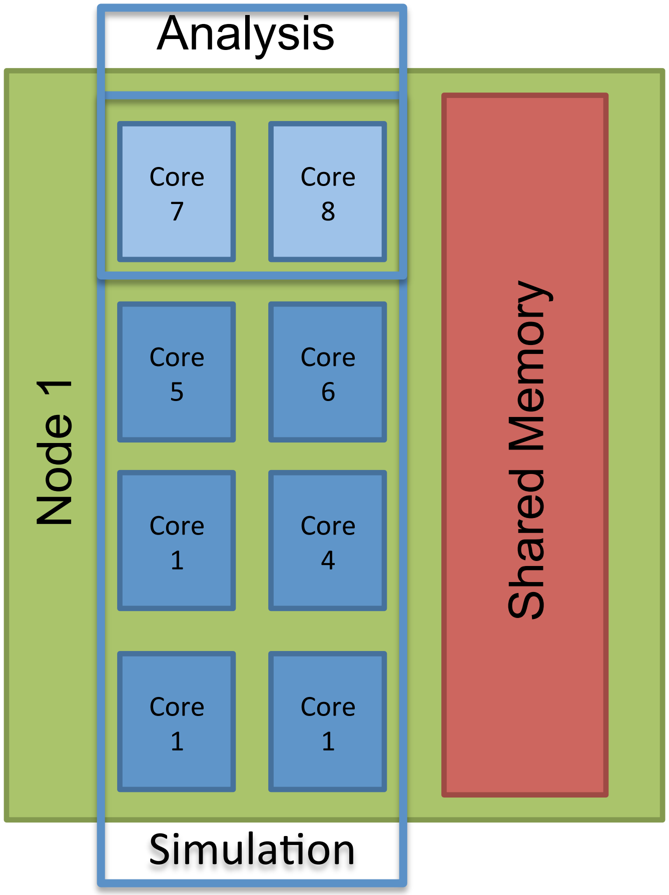
few have seen real machines with it,

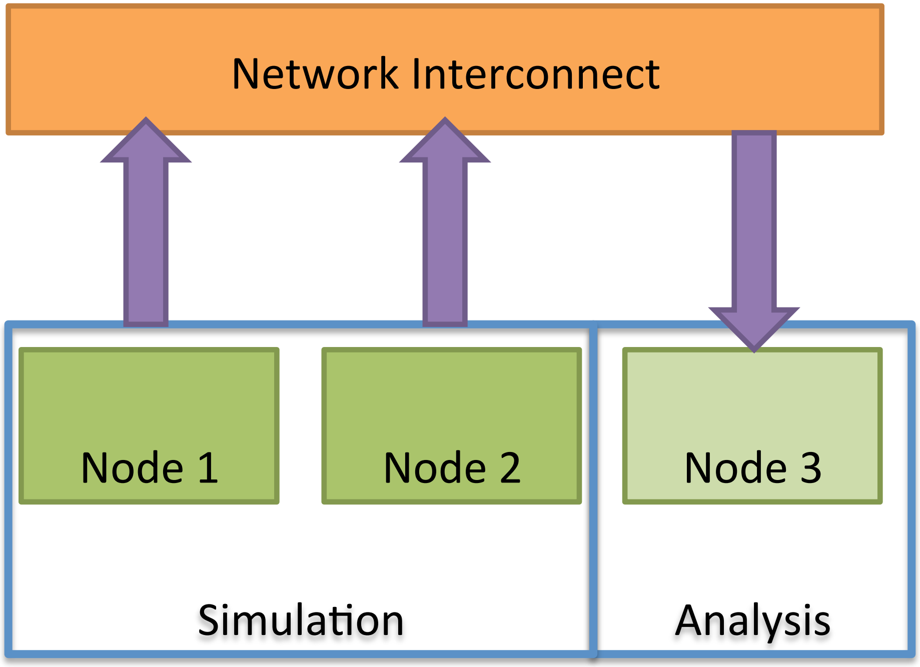
even fewer have ran applications on those machines …

Remaining Challenges

- Burst Buffers are **NOT** the magic I/O silver bullet
  - I/O contention still a problem if we exceed the burst buffer capability
  - Burst buffers improve offloading bandwidth but do **NOT** help uploading data from storage for analysis and visualization

In-situ and *In-transit* Analysis





Example of tools:

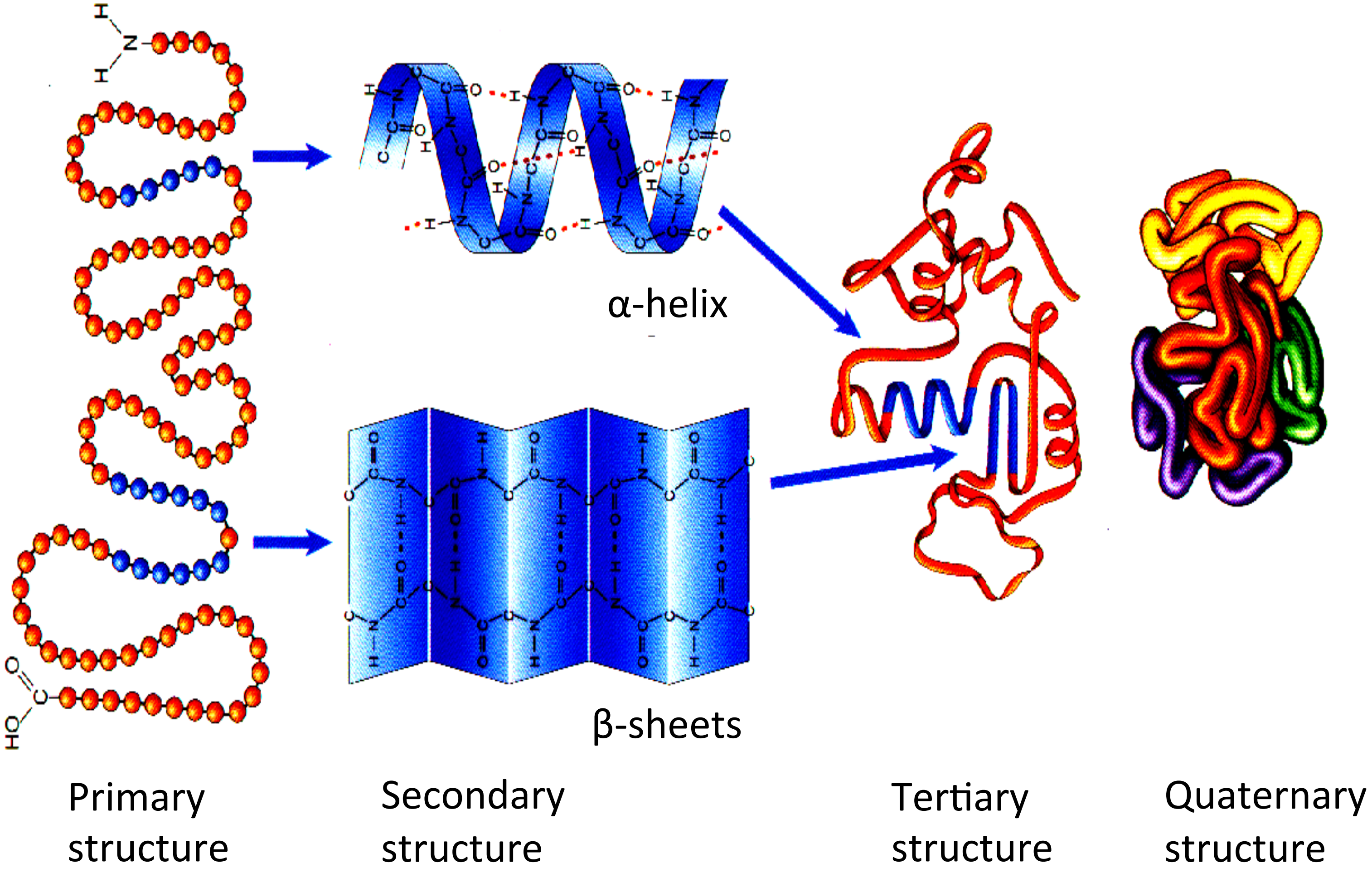
- DataSpaces (Rutgers U.)
- DataStager (GeorgiaTech)

In-situ and *In-transit* MD Analysis

"***In-Situ Data Analysis and Indexing of Protein Trajectories***," Travis Johnston, Buyu Zhang, Adam Liwo, Silvia Crivelli, and Michela Taufer. JCC 2017.

- Integrate *in-situ* and *in-transit* analysis in MD simulations
  - Mitigate data movements by moving the analysis close to the simulation
- Define *in-situ* algorithms for runtime analysis of molecular structures in molecular dynamics simulations
  - Capture conformation changes of individual secondary structures
  - Capture changes of tertiary structures (i.e., a secondary structure with respect to other secondary structures)

Protein Structures



α-helix

β-sheets

Primary structure

Secondary structure

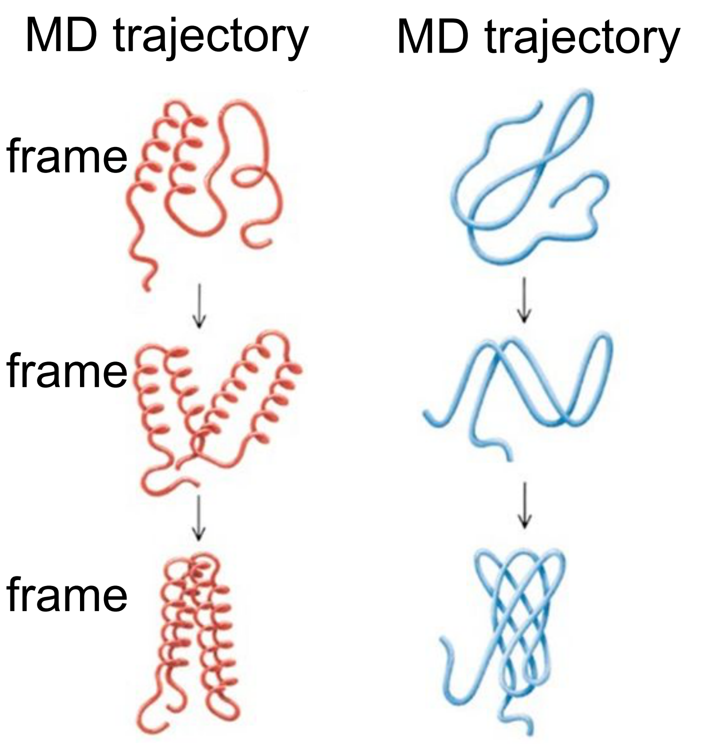
Tertiary structure

Quaternary structure

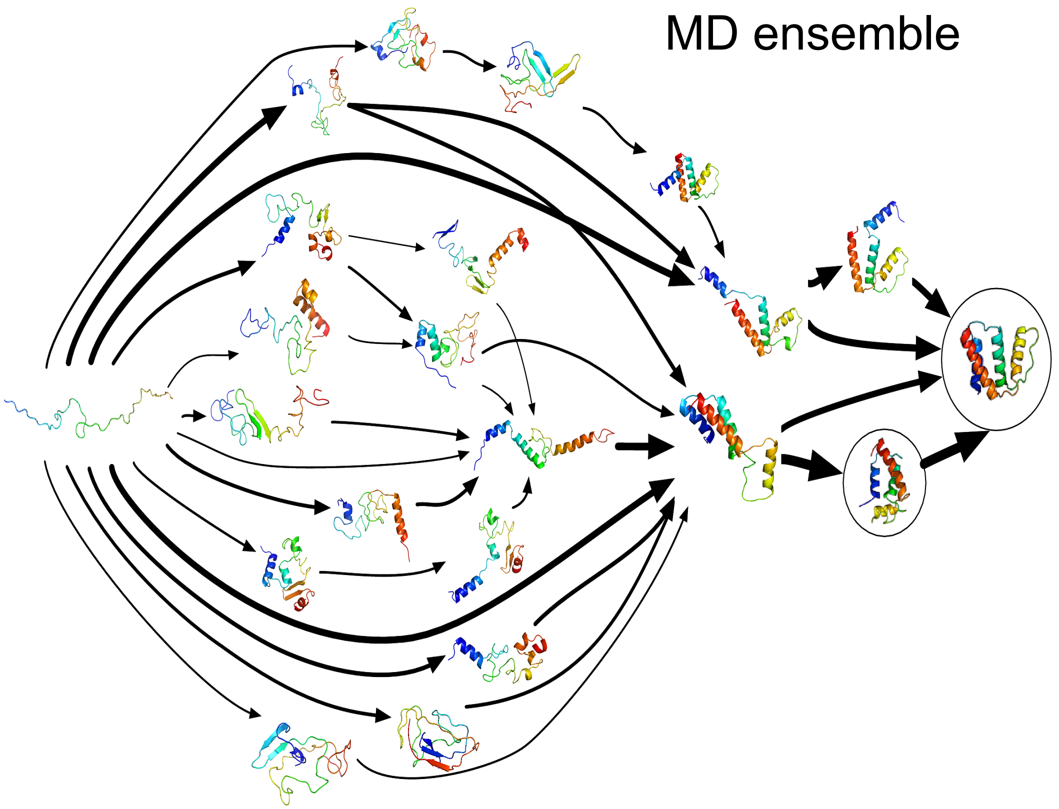
From: http://schoolworkhelper.net/protein-structures-primary-secondary-tertiary-quaternary/

Molecular Dynamics Simulations

Use computer simulations to study the physical movements of atoms and molecules



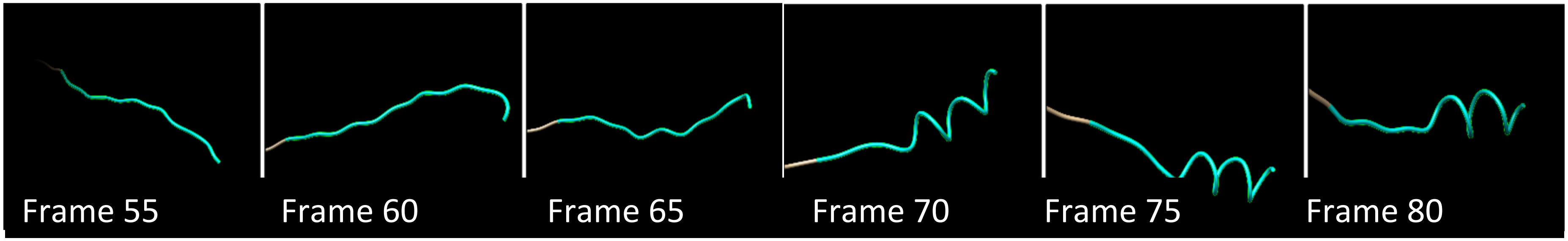
From: http://images.slideplayer.com/18/5667960/slides/slide_1.jpg



By Vincent Voelz - Sent to the uploader personally, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=218912120

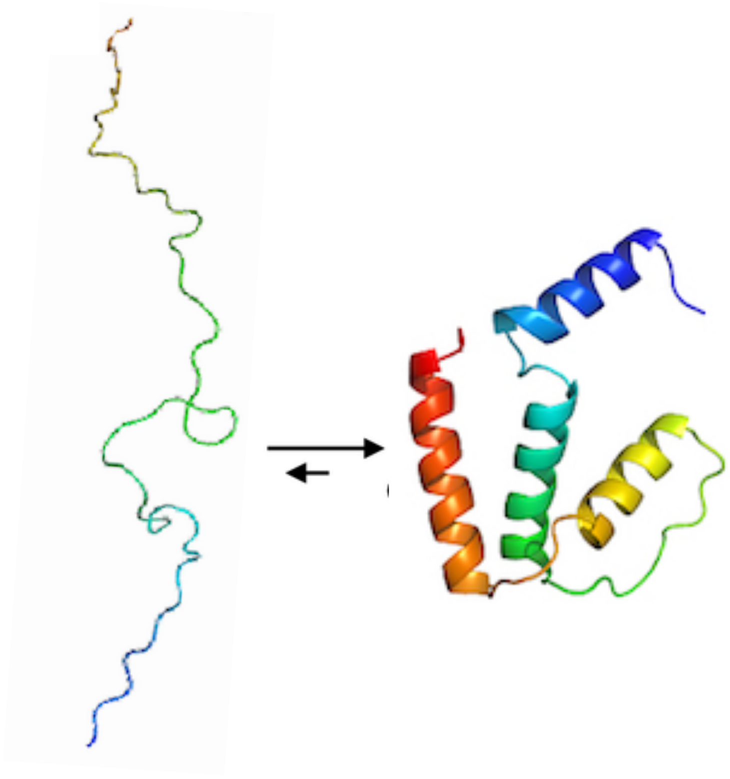
Analysis Requirements

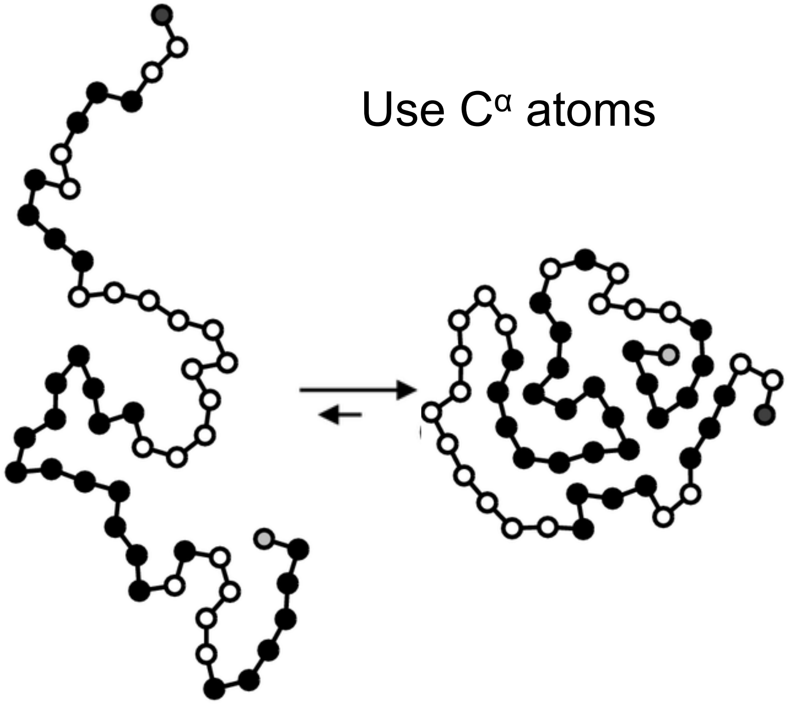
Frames of an MD trajectory:



- We want to capture what is going on in each frame **without**:
  - Disrupting the simulation (e.g., stealing CPU and memory on the node)
  - Moving all the frames to a central file system and analyzing them once the simulation is over
  - Comparing each frame with past frames of the same job
  - Comparing each frame with frames of other jobs

Modeling Molecules



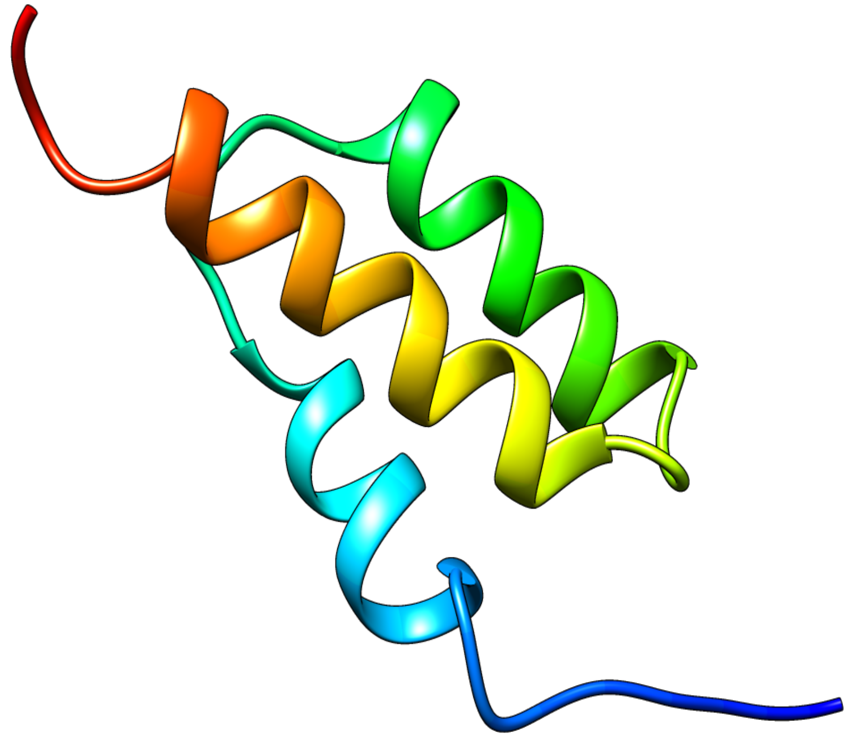


Use C$^{\alpha}$ atoms

From: By Tomixdf (talk) - Own work (Original text: self-made), Public Domain, https://commons.wikimedia.org/w/index.php?curid=23662306

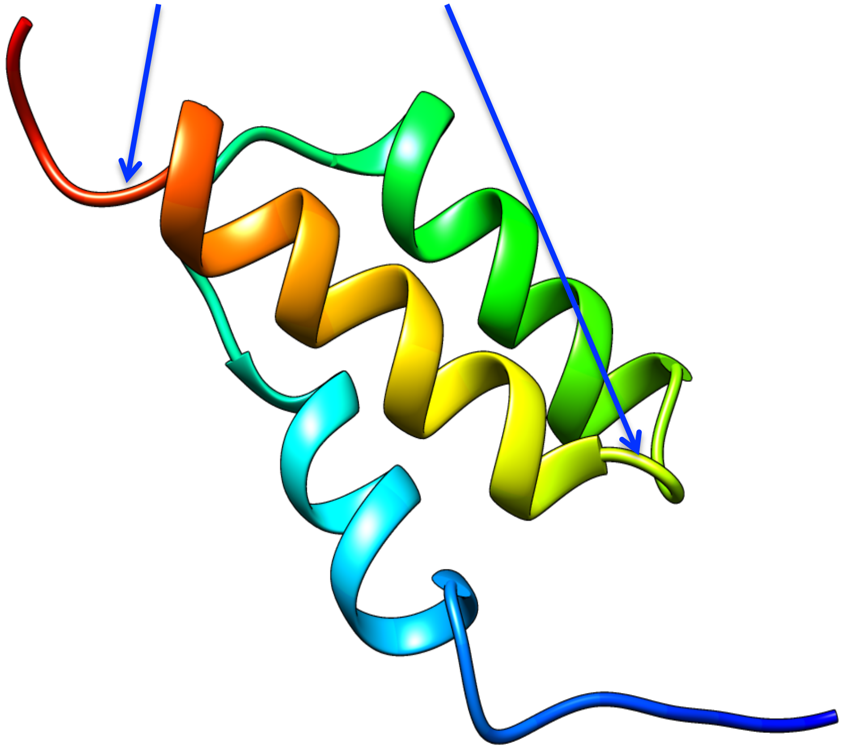
Capturing Secondary Structures

Given a frame of an
MD job at time *t*



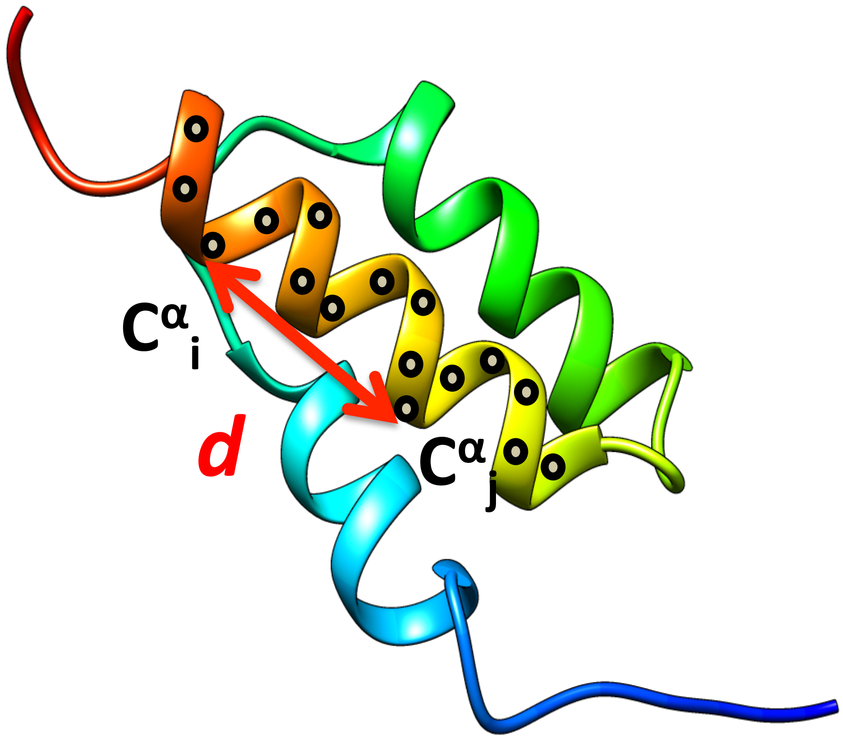
Capturing Secondary Structures

Define the substructure:
start and **stop** amino acids



Capturing Secondary Structures

Measure the distance between C^{α}_{j} and C^{α}_{i}



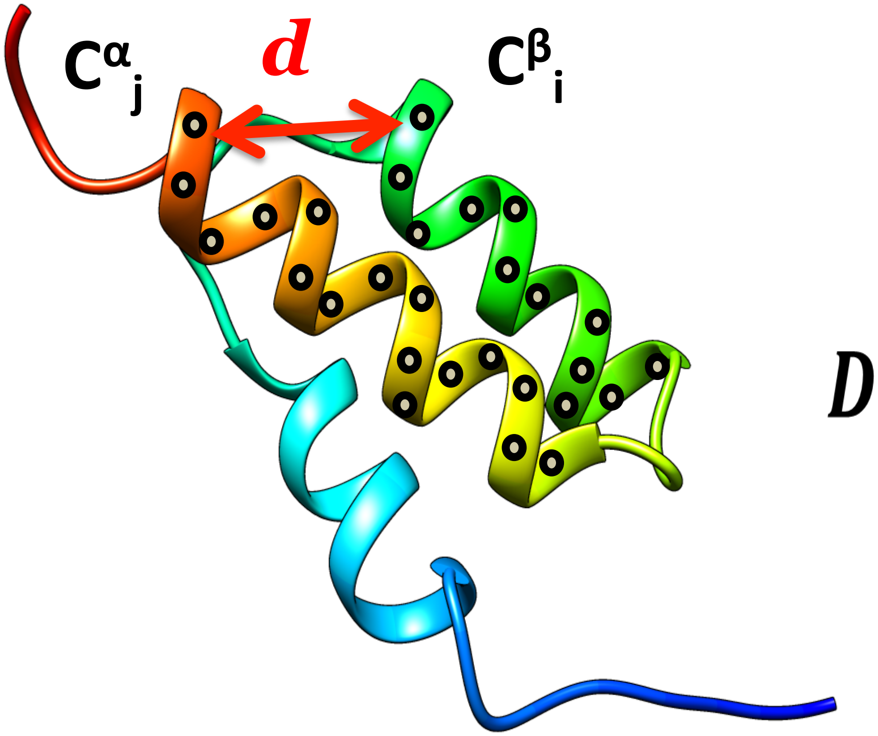
Build the **substructure Euclidean Distance Matrix (D)**

$$D = \begin{array}{ccccccc} & & & C^{\alpha}_{i} & & & \\ & 0 & \times & \times & \times & \times & \times \\ C^{\alpha}_{j} & \times & 0 & d & \times & \times & \times \\ & \times & d & 0 & \times & \times & \times \\ & \times & \times & \times & 0 & \times & \times \\ & \times & \times & \times & \times & 0 & \times \\ & \times & \times & \times & \times & \times & 0 \end{array}$$

Compute largest eigenvalue → λ_{max}

Capturing Tertiary Structures

Measure the distance between C^{α}_{j} and C^{β}_{i}



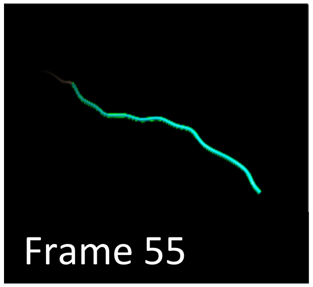
Build a **bipartite distance matrix** by comparing two substructures

$$D = \begin{array}{c|cccccc} & & & & i & & \\ \hline & 0 & 0 & 0 & X & X & X \\ j & 0 & 0 & 0 & d & X & X \\ & 0 & 0 & 0 & X & X & X \\ & X & d & X & 0 & 0 & 0 \\ & X & X & X & 0 & 0 & 0 \\ & X & X & X & 0 & 0 & 0 \end{array}$$

Compute largest eigenvalue → λ_{max}

Proxy for Conformations' Changes

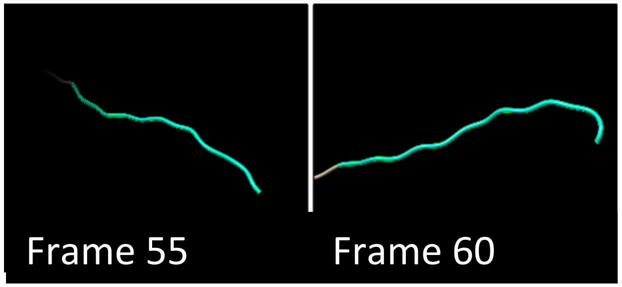
Frames of an MD job:



$\boldsymbol{\lambda_{55}}$

Proxy for Conformations' Changes

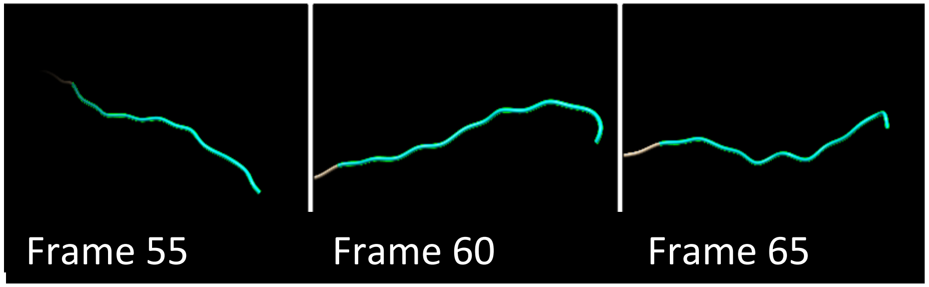
Frames of an MD job:



$\boldsymbol{\lambda}_{55}$ $\boldsymbol{\lambda}_{60}$

Proxy for Conformations' Changes

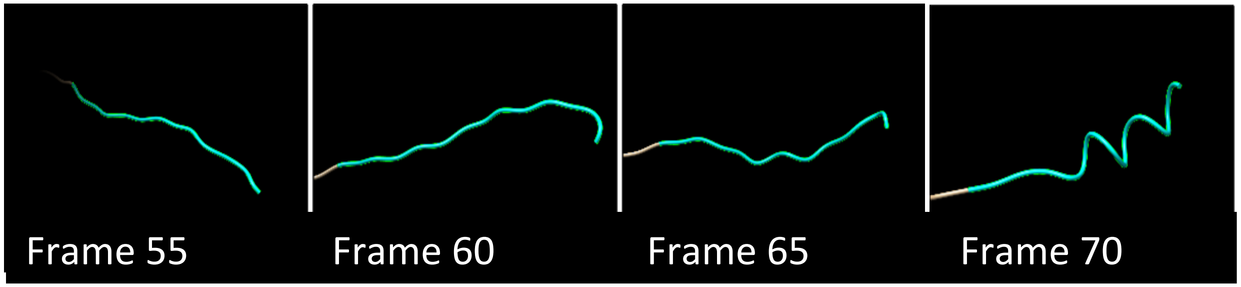
Frames of an MD job:



λ_{55} λ_{60} λ_{65}

Proxy for Conformations' Changes

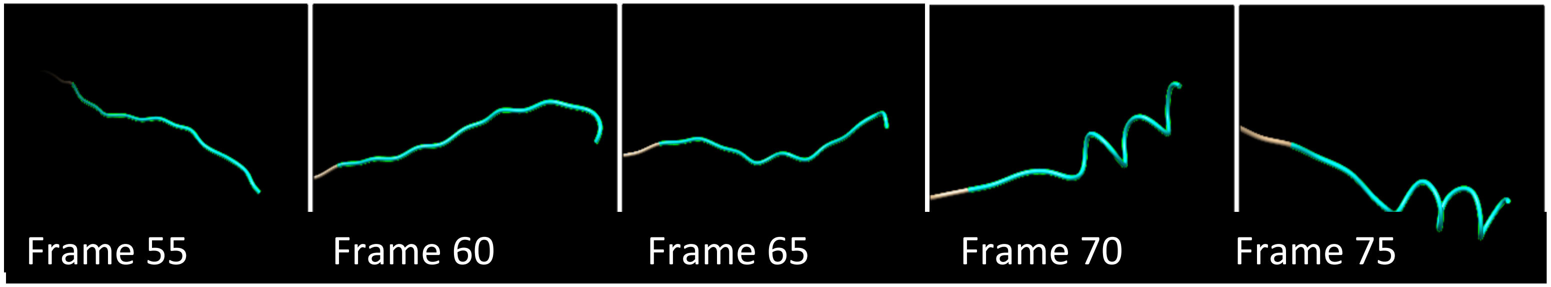
Frames of an MD job:



λ_{55} λ_{60} λ_{65} λ_{70}

Proxy for Conformations' Changes

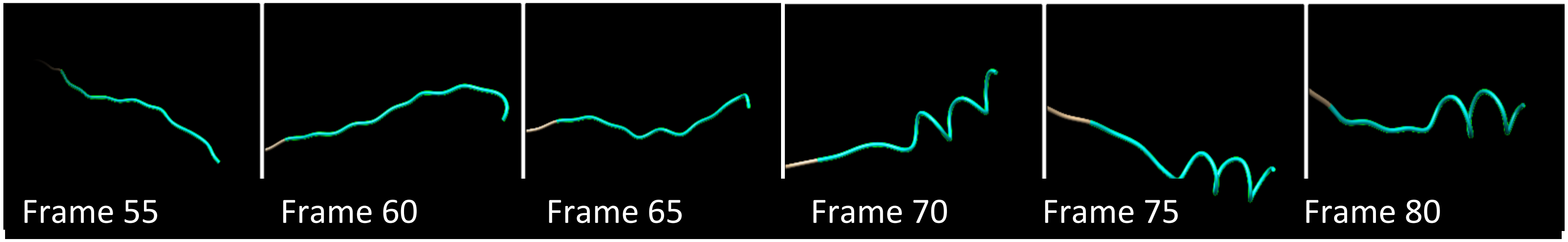
Frames of an MD job:



λ_{55} λ_{60} λ_{65} λ_{70} λ_{75}

Proxy for Conformations' Changes

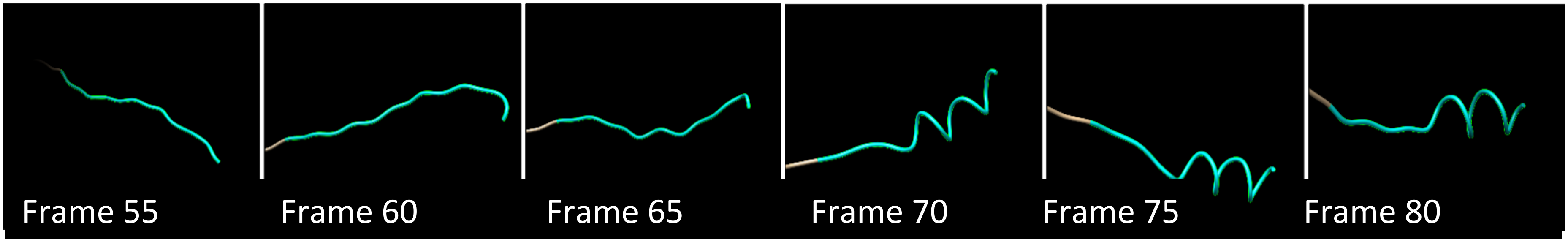
Frames of an MD job:



λ_{55} λ_{60} λ_{65} λ_{70} λ_{75} λ_{85}

Proxy for Conformations' Changes

Frames of an MD job:



λ_{55} λ_{60} λ_{65} λ_{70} λ_{75} λ_{85}

Distance between two max eigenvalues serves as a proxy for distance between the two associated conformations

Proxy for Conformations' Changes

Distance between two max eigenvalues serves as a proxy for distance between the two associated conformations

- Euclidean distance matrix D is symmetric
- Eigenvalues of symmetric, real matrices are stable
  - Small perturbations of D result in only small changes in the eigenvalues
  - Euclidean distance matrix is insensitive to rigid transformation

$$\begin{array}{cc}
 & \alpha\text{-carbon} \\
\alpha\text{-carbon} & \begin{bmatrix}
0 & x & x & x & x & x & x & x & x & x & x \\
x & 0 & x & x & x & x & x & x & x & x & x \\
x & x & 0 & x & x & x & x & x & x & x & x \\
x & x & x & 0 & x & x & x & x & x & x & x \\
x & x & x & x & 0 & x & x & x & x & x & x \\
x & x & x & x & x & 0 & x & x & x & x & x \\
x & x & x & x & x & x & 0 & x & x & x & x \\
x & x & x & x & x & x & x & 0 & x & x & x \\
x & x & x & x & x & x & x & x & 0 & x & x \\
x & x & x & x & x & x & x & x & x & 0 & x \\
x & x & x & x & x & x & x & x & x & x & 0
\end{bmatrix}
\end{array}$$

- Use only largest eigenvalue in distance matrix

$\lambda max = \lambda1 < \lambda2 < \lambda3 < \lambda4 < \lambda5 = \lambda min$

$\lambda1 + \lambda2 + \lambda3 + \lambda4 + \lambda5 = 0$

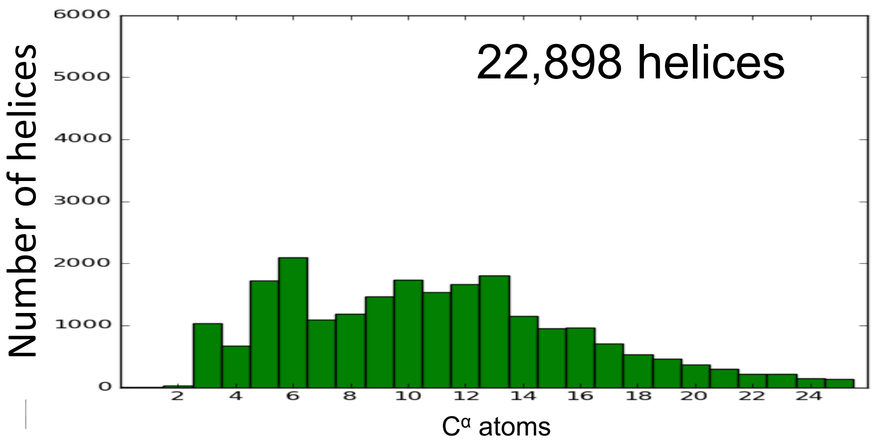
$\lambda1 \gg \lambda2 \sim \lambda3 \sim \lambda4 \sim 0$

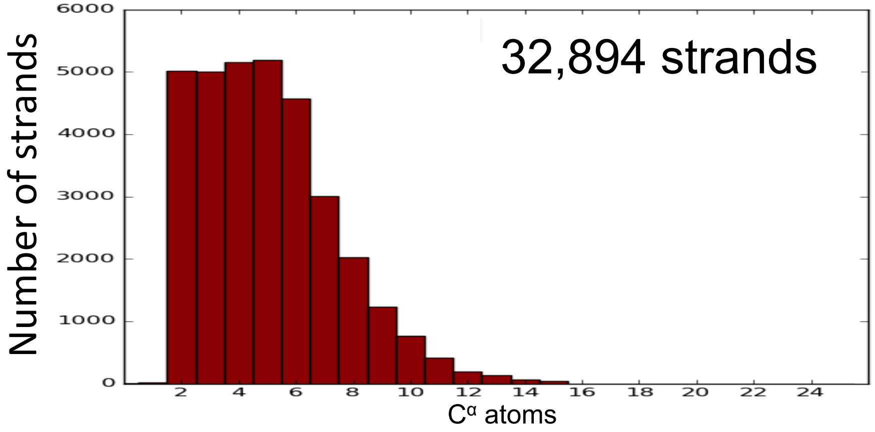
$\lambda max = \lambda1 \sim - \lambda5 = - \lambda min$

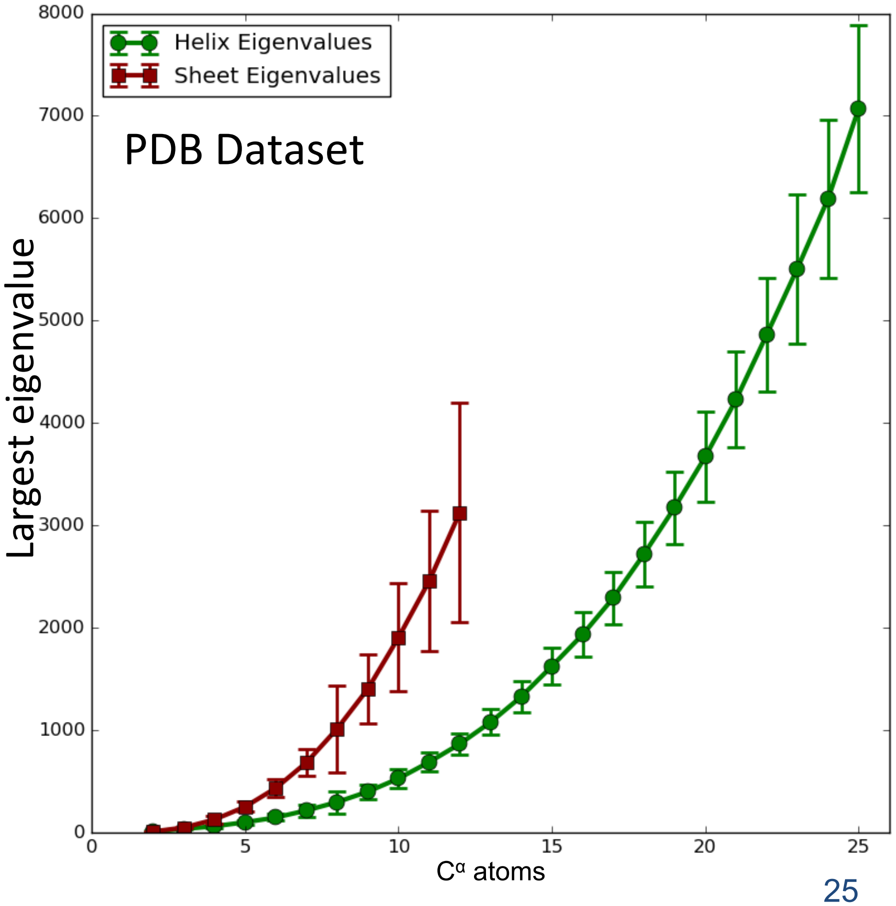
"***In-Situ Data Analysis and Indexing of Protein Trajectories***," Travis Johnston, Buyu Zhang, Adam Liwo, Silvia Crivelli, and Michela Taufer. JCC 2017.

Mapping Largest Eigenvalues to Structures

PDB dataset: 3,197 different proteins including 22,898 helices and 32,894 strands

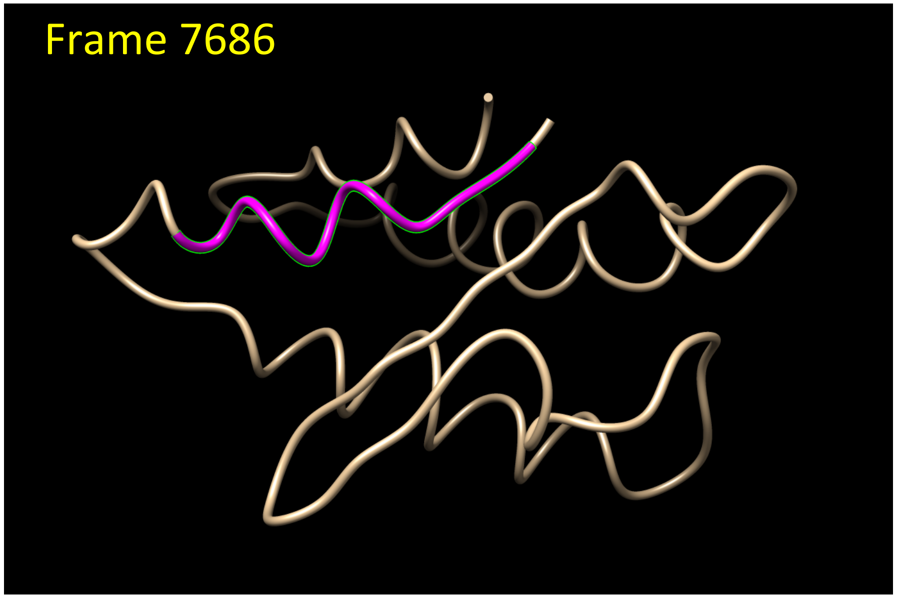


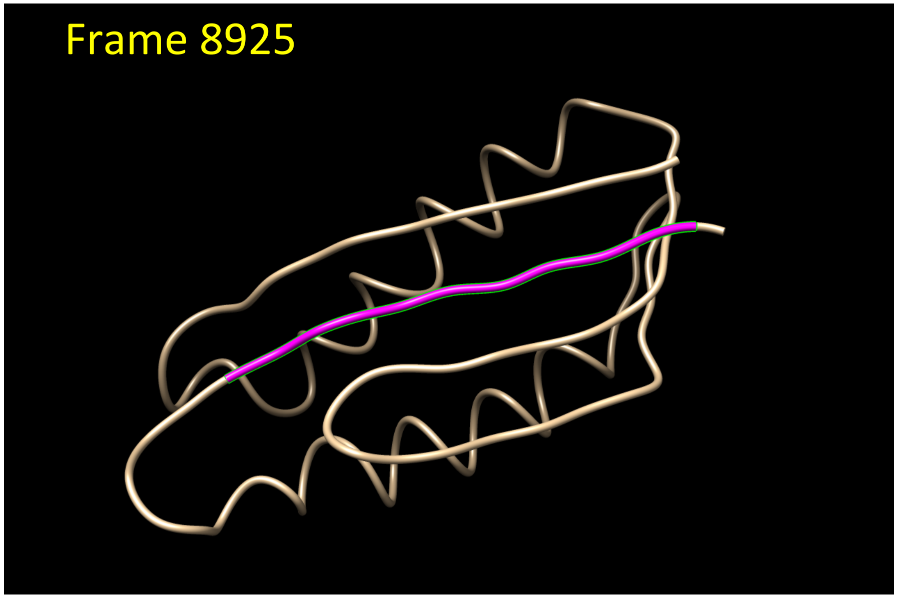




Case Study I: 2MQ8 Protein

- Canonical simulation of 2MQ8 protein including both α helices and β strands
  - After ~9M steps α helices pack tighter and change into β strands

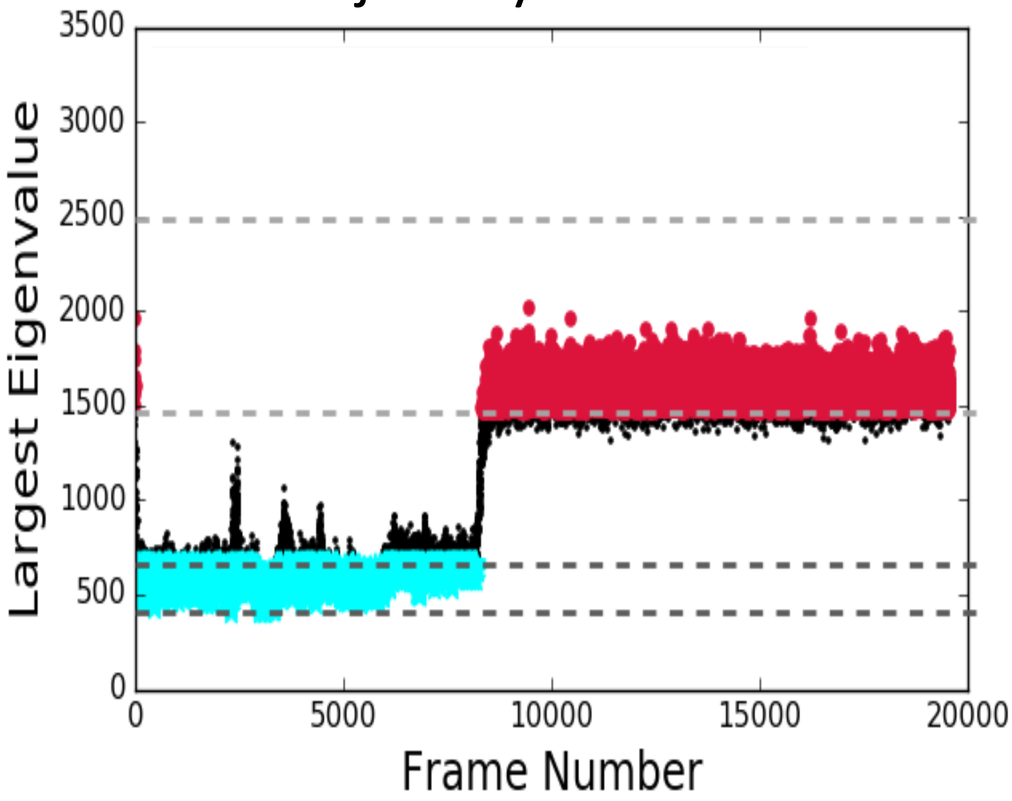




Can the eigenvalue analysis capture the conformational change?

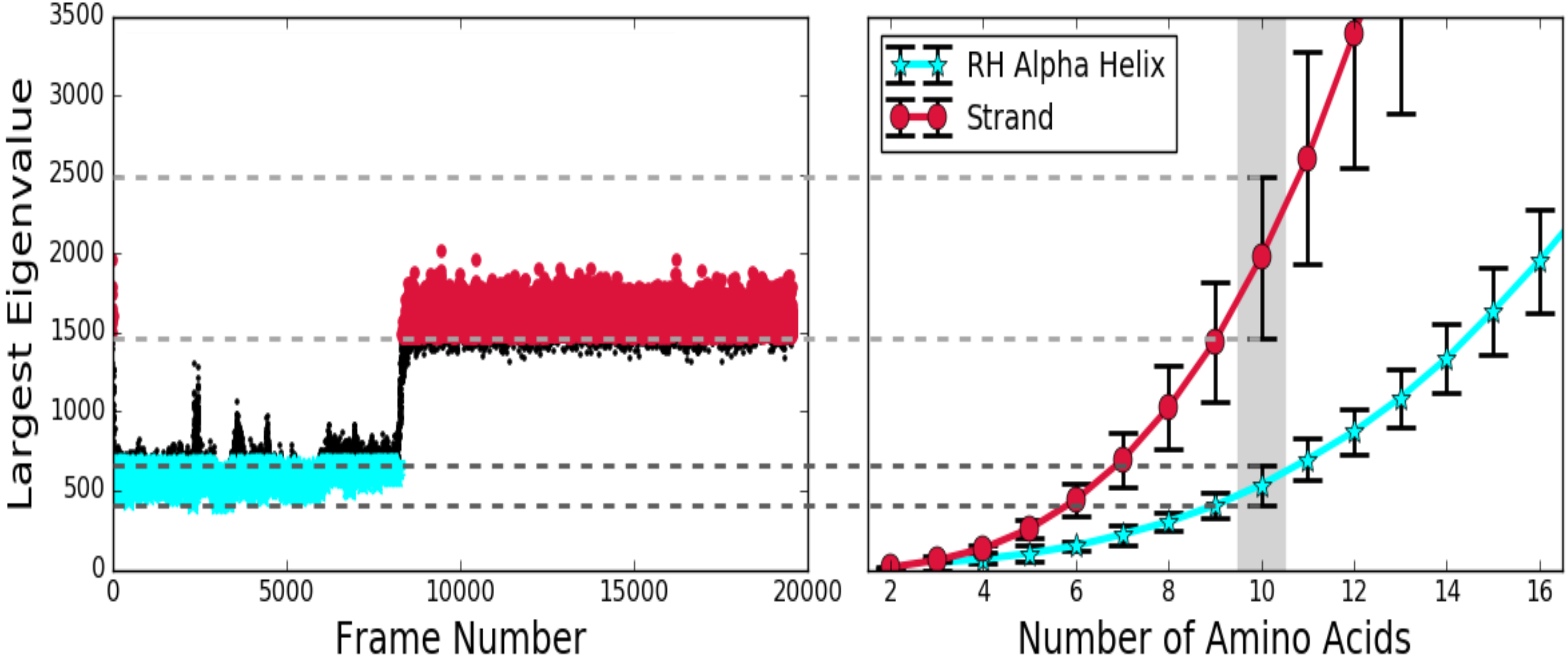
Case Study I: 2MQ8 Protein

Compute largest eigenvalue of 3rd strand (10 amino acids) for each trajectory frame



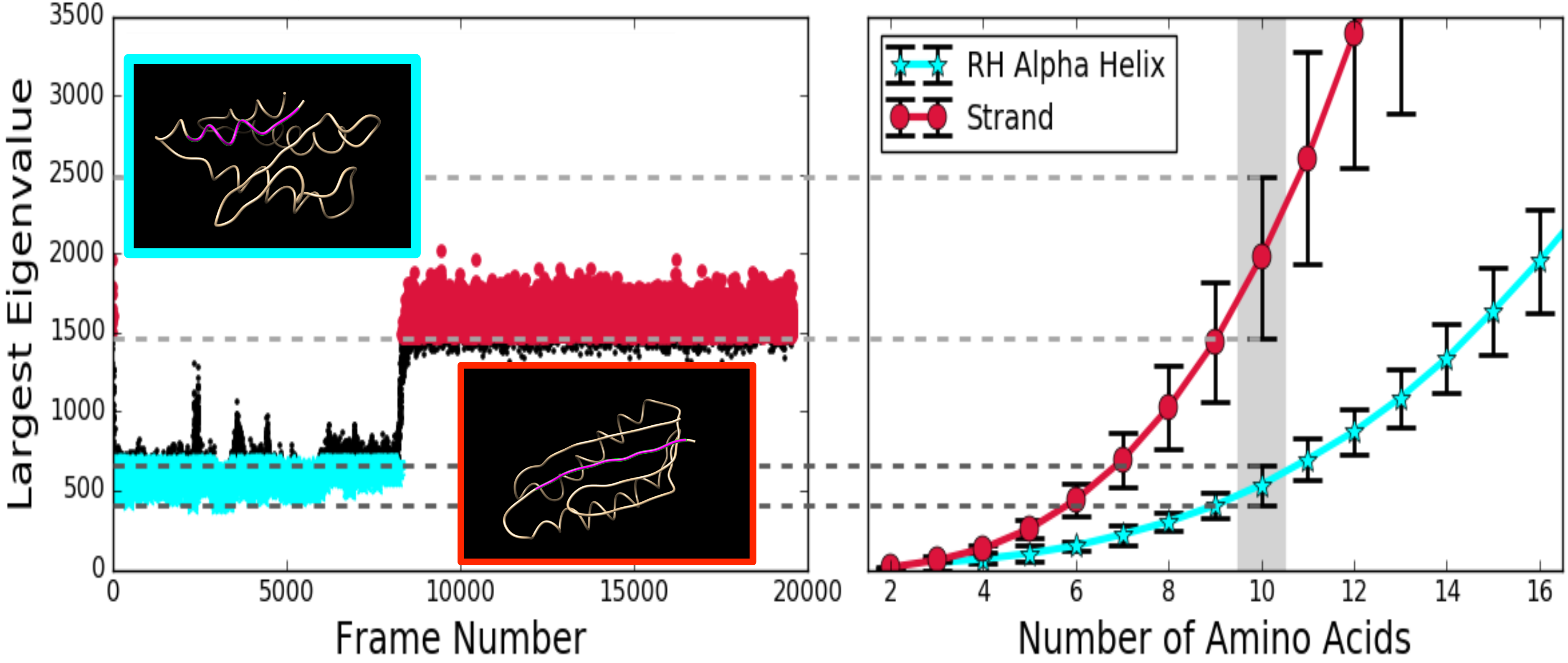
Case Study I: 2MQ8 Protein

Compute largest eigenvalue of 3rd strand (10 amino acids) for each trajectory frame



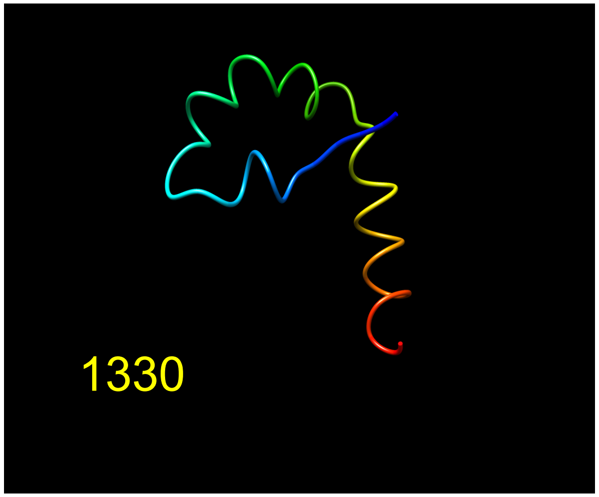
Case Study I: 2MQ8 Protein

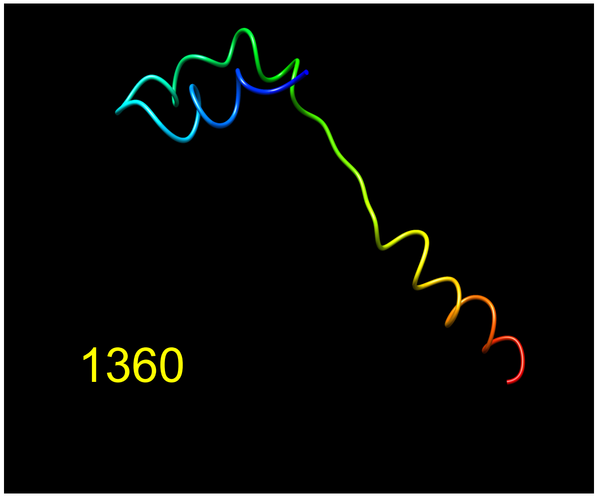
Compute largest eigenvalue of 3rd strand (10 amino acids) for each trajectory frame

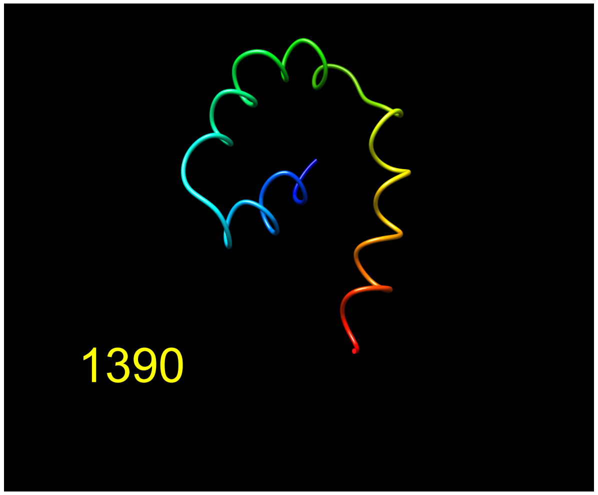


Case Study II: Capturing Movement of α-helices

Capture movement of structures with respect to each other

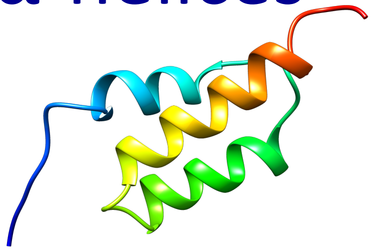




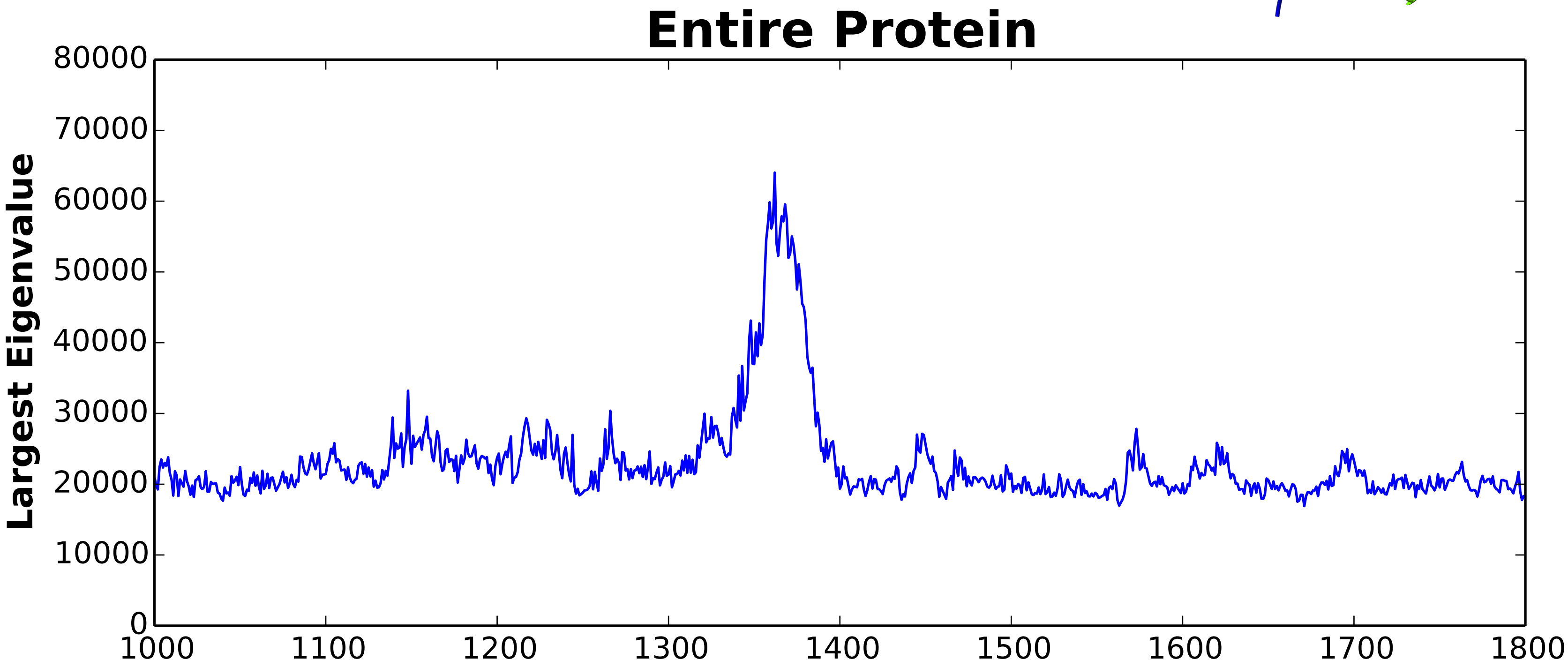


Can the eigenvalue analysis capture the movement of helices ?

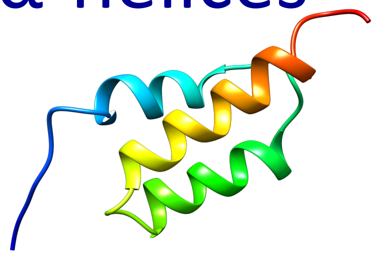
Case Study II: Capturing Movement of α-helices



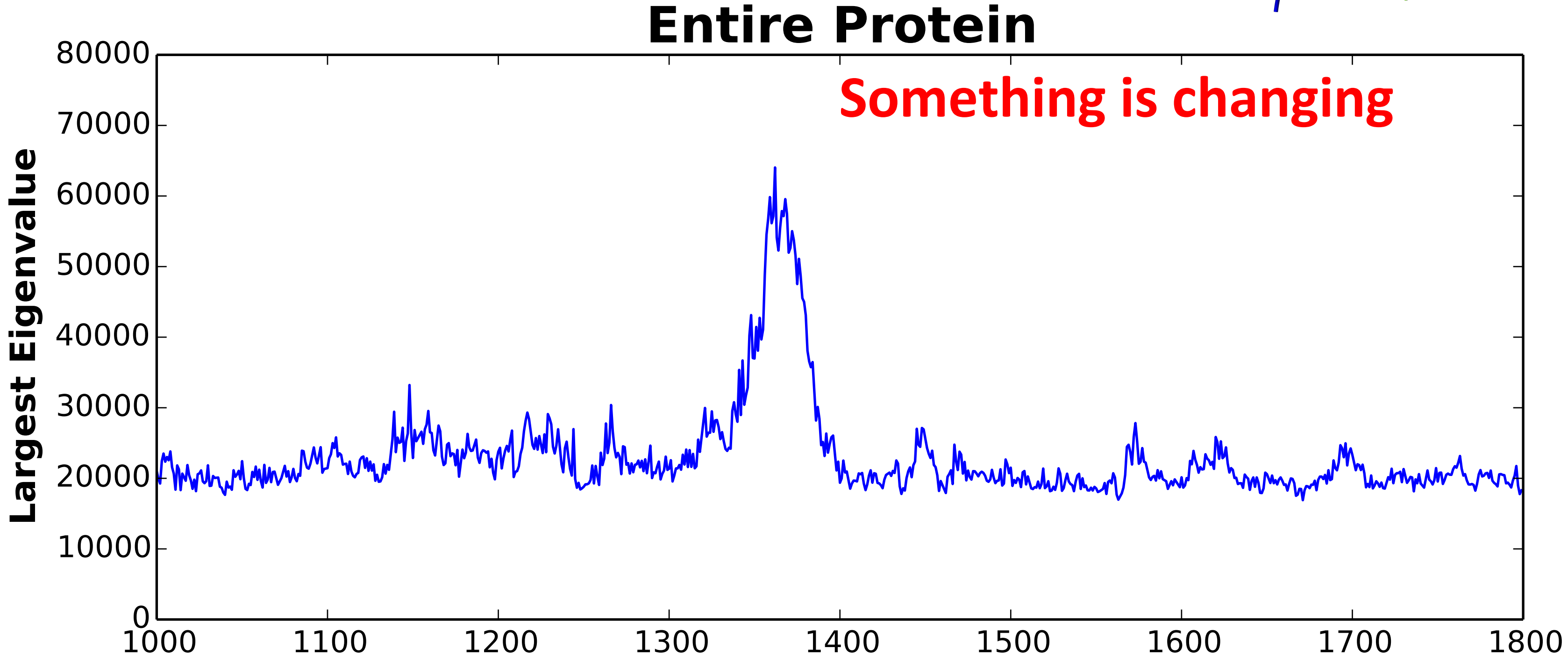
Monitor largest eigenvalue of entire protein



Case Study II: Capturing Movement of α-helices

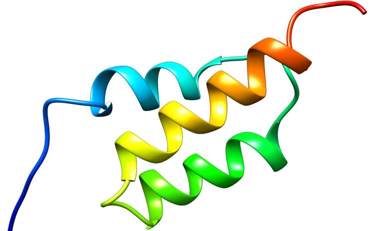


Monitor largest eigenvalue of entire protein

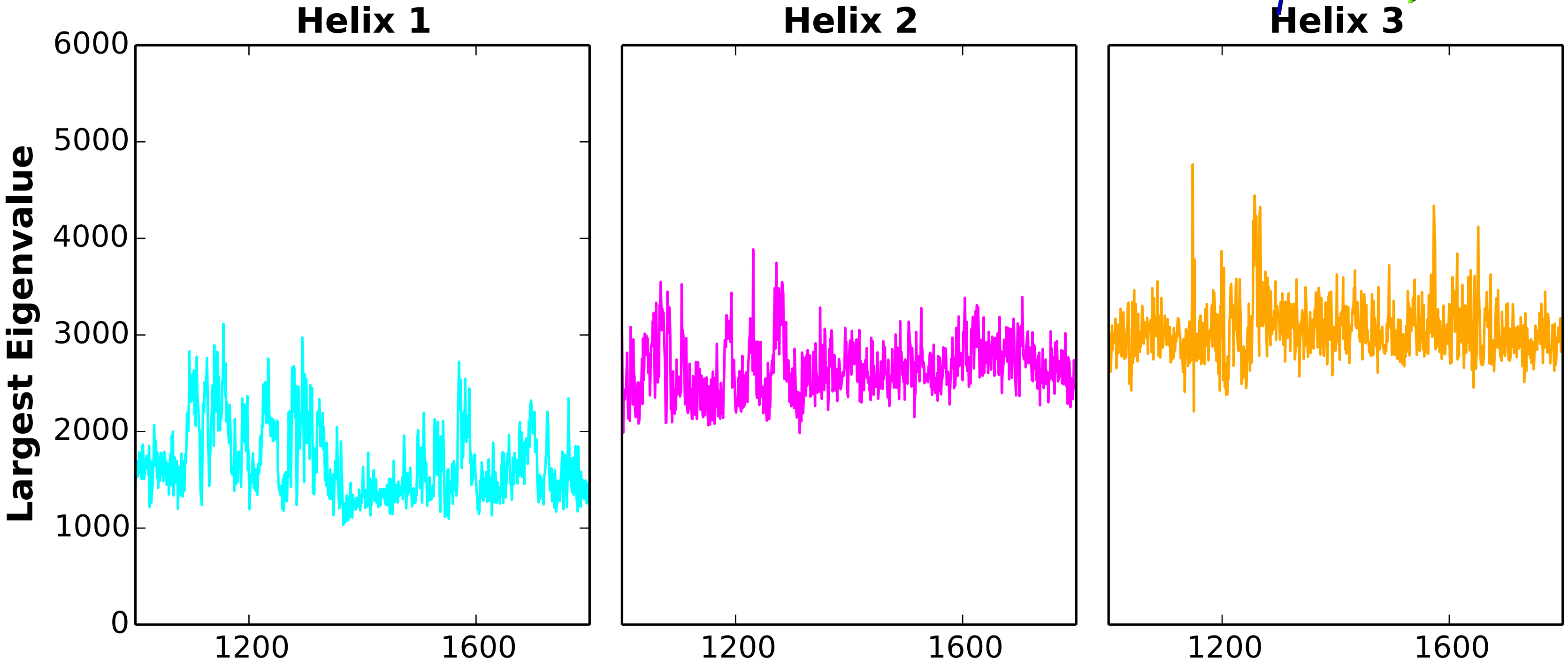


Something is changing

Case Study II: Capturing Movement of α-helices

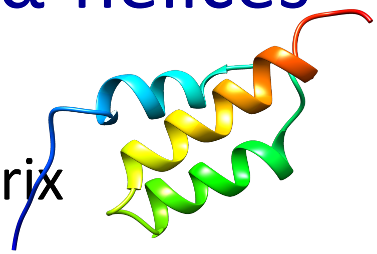


Monitor largest eigenvalue of single helices

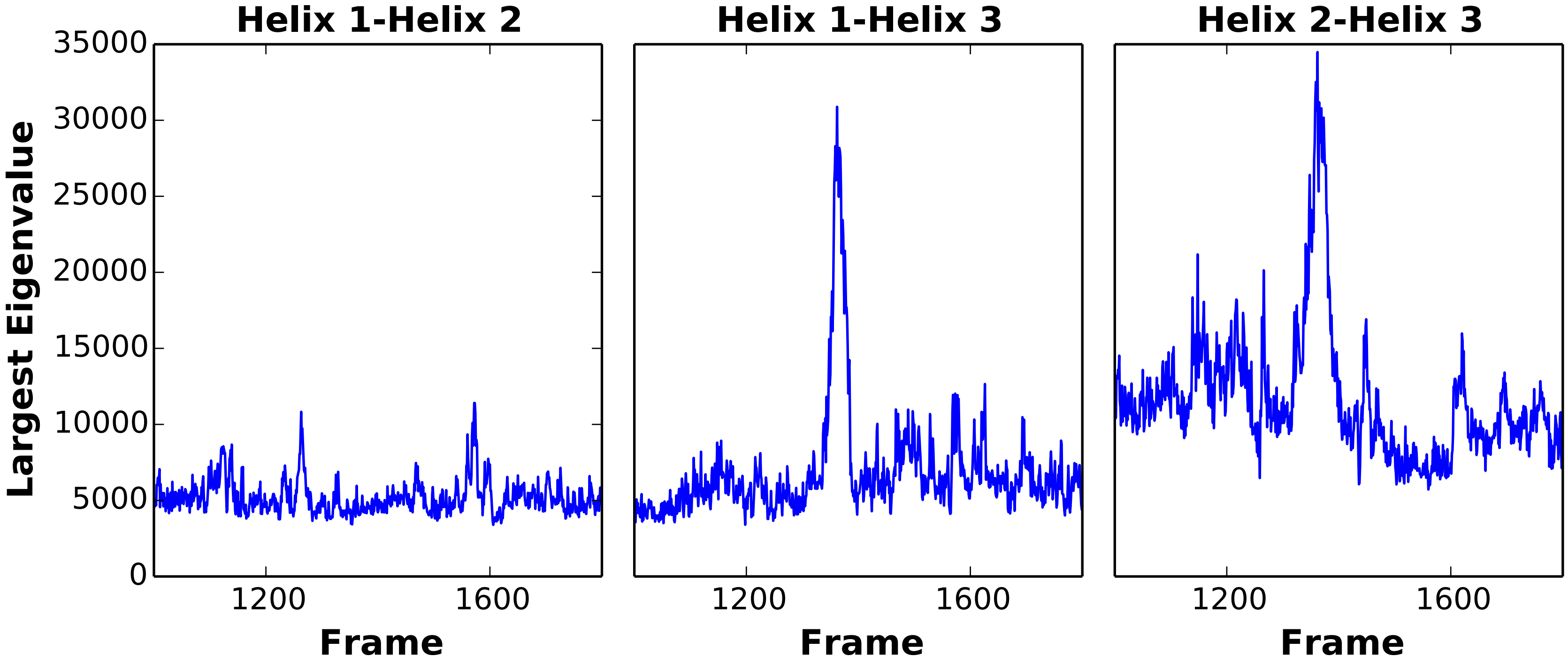


Individual α-helices (Helix 1, Helix 2, and Helix 3) appear stable

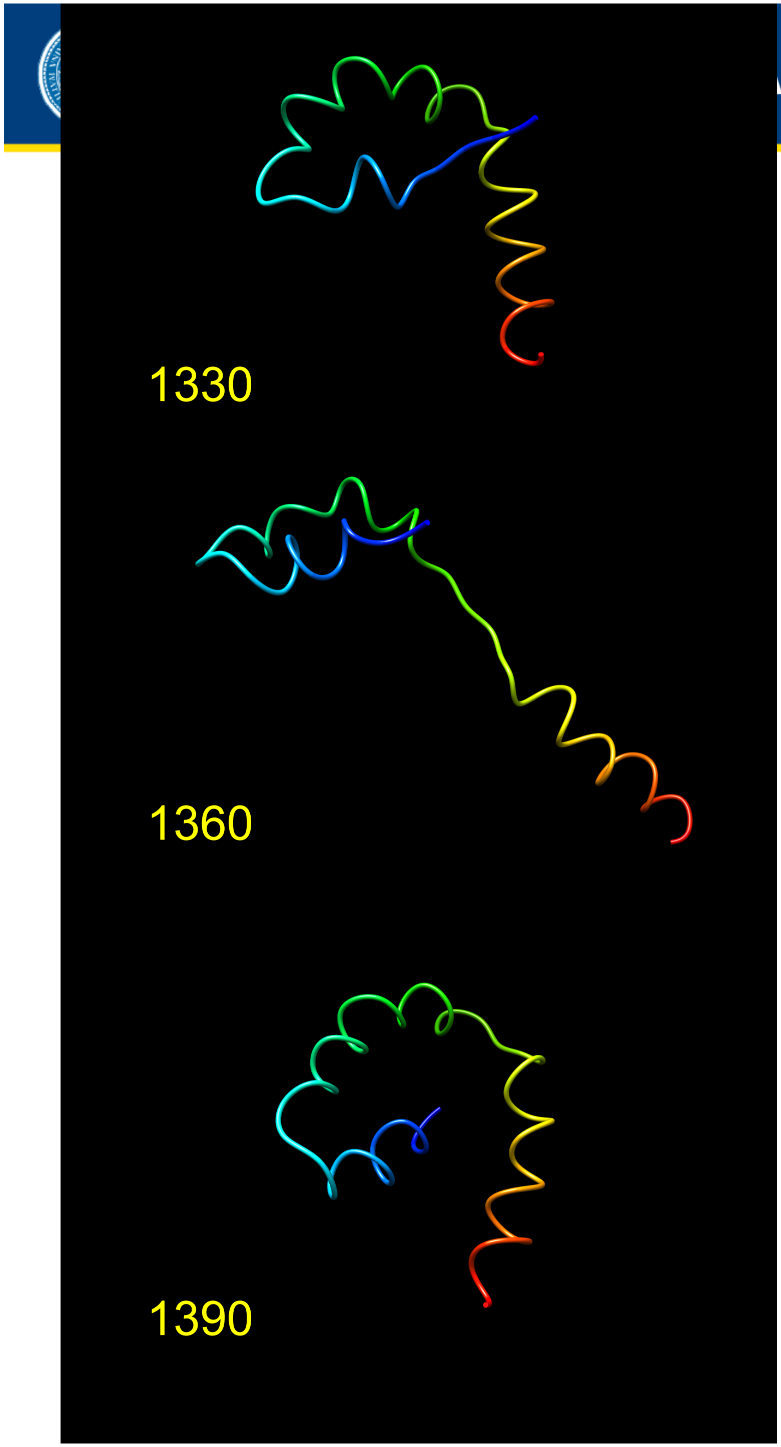
Case Study II: Capturing Movement of α-helices



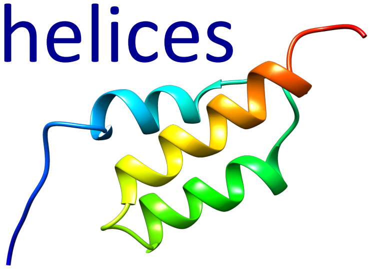
Monitor largest eigenvalue of bipartite distance matrix

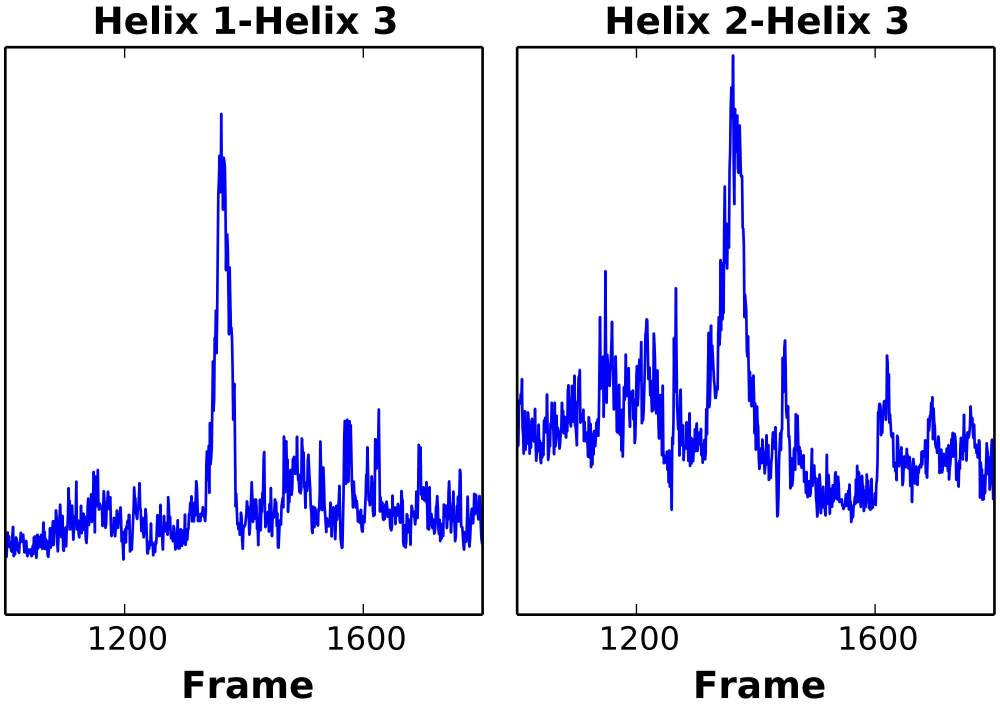


First and second α-helices appear stable; third helix moves



Case Study II: Capturing Movement of α-helices





Large relative change between two pairs of α-helices

Perspectives

- The traditional HPC point of view:

" [...] the results would not be of interest to the Supercomputing community given that the method has not been parallelized, does not discuss how the method might be parallelized, etcetera. Hence, the work may be more appropriate for a computational chemistry journal or the IEEE Big Data Conference."

Anonymous Feedback

10th IEEE International Scalable Computing Challenge SCALE 2017

- Co-located with IEEE/ACM CCGrid, Madrid, Spain, May 14-17, 2017
- Objectives:
  - Showcase real-world problems solved using computing that **scales**
  - Advances in applications development and infrastructure to enable scaling
- Submissions:
  - Format: White papers (6-page max. in IEEE format)
  - Deadline: **Jan. 31, 2017**
  - All selected contributions in conf. proceedings
- Awards:
  - First Place: $1,000
  - Second Place: $500

- Chairs: Michela Taufer (U Delaware) & Antonio J. Peña (BSC)

https://www.arcos.inf.uc3m.es/wp/ccgrid2017/calls/scale-challenge