Enabling Scalable Data Analysis of Computatioal Structural Biology Datasets on Distributed Memory Systems Supported by the MapReduce Paradigm

Boyu Zhang (PhD Student), Michela Taufer (Advisor)
University of Delaware

Thesis Statement

- A distributed classification and clustering of structural biology datasets is feasible and scalable
- Our approach uses the MapReduce programming model to transform relevant data properties into metadata concurrently and to extract science from metadata
- We assess out approach on three structural biology datasets and relevant scientific problems

Dataset	Scientific problem	Metadata	Analysis
2D: ACGUGCCACGAU…	Classification of RNA secondary structures	Strings representing chunk-based secondary structures …(((….)))…	Classification: statistical-based
3D: x_1, y_1, z_1, x_2, y_2, z_2, …	Clustering of ligand geometries	3D point representing geometric shape features	Clustering: octree-based
4D: x_{t1},y_{t1}, z_{t1}, x_{t2}, y_{t2}, …	Clustering of protein-folding trajectories	A set of 3D points representing geometric shape features in time	Clustering: hierarchical probabilistic-based

RNA Secondary Structures

Scientific Problem

Enable scalable secondary structure predictions of long RNA sequences

acGUGCCACGauucaaCGUGGCACag

Hairpin structure

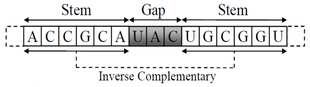
uCGACUGuAAAaaaGCGGgGCGACUUUCAG
UCGcucuuuuuGUCGCgCGC

Pseudoknot structure

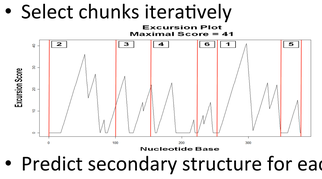
Method

From RNA sequence to metadata

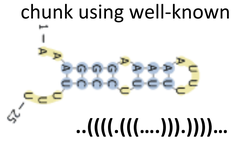
- Identify inversions in the sequence



- Build inversion excursions
- Select chunks iteratively



- Predict secondary structure for each chunk using well-known programs



..((((.(((....))).))))...

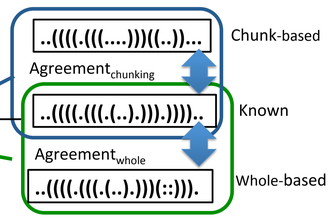
From metadata to scientific knowledge

- Reconstruct the entire secondary structure by concatenate chunked structures

.....(((((((.........))))))).. + ...((((..(((..))))....((....).)))...

.....(((((((.........)))))))..........((((..(((..)))..... ((...)).))).

- Access likelihood vs. known structure and non-chunked (whole) approach

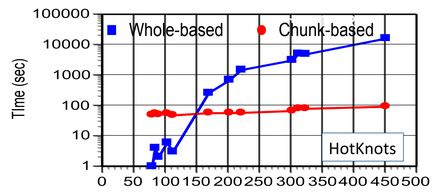


AR = Agreement$_{chunking}$ / Agreement$_{whole}$

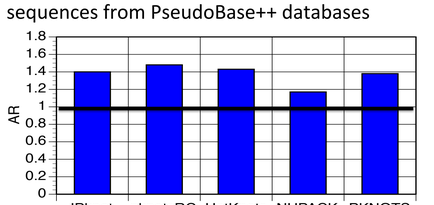
AR > 1: predictions with our method are more accurate

Results

Execution time of our chunk-based method vs. traditional method on 8 Hadoop nodes



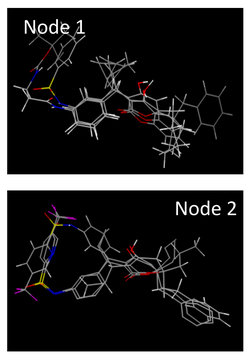
Accuracy ratio (AR) for 23 pseudoknotted sequences from PseudoBase++ databases



Ligand Geometries

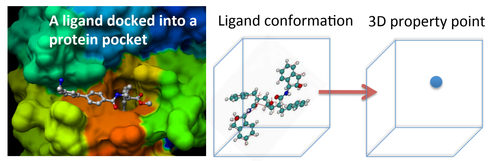
Scientific Problem

Enable comparison of ligand conformations and identify predominate conformations across distributed disks

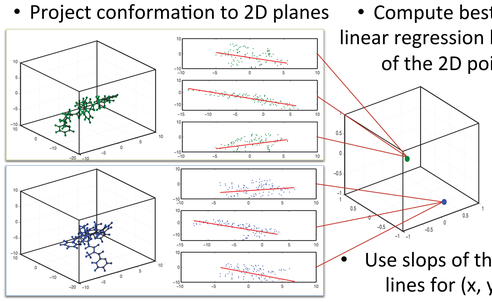


Method

From ligand conformation to metadata



- Project conformation to 2D planes
- Compute best-fit linear regression line of the 2D points

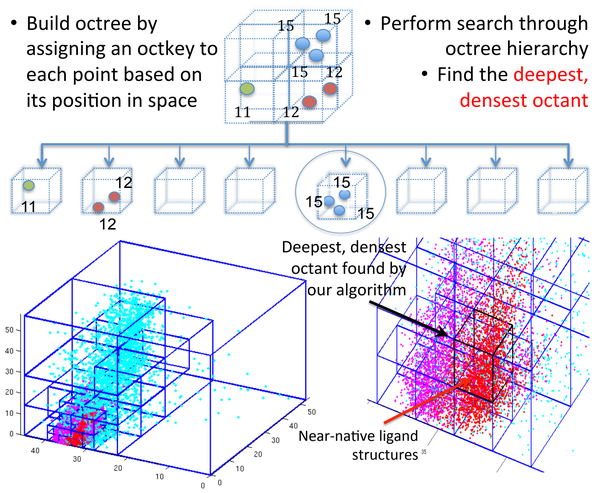


- Use slops of three lines for (x, y, z)

Similar conformations mapped to closer 3D points in space

From metadata to scientific knowledge

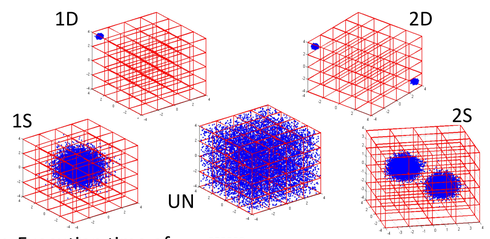
- Build octree by assigning an octkey to each point based on its position in space



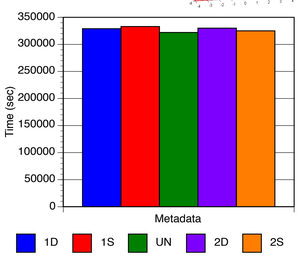
- Perform search through octree hierarchy
- Find the deepest, densest octant

Results

Five datasets with different metadata clusters



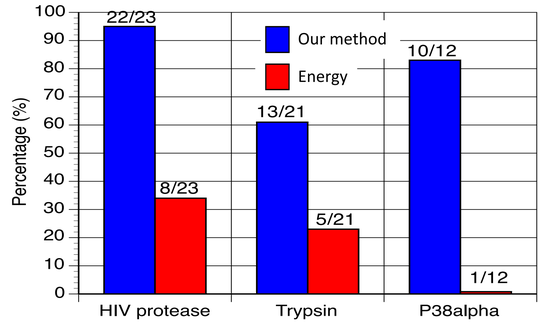
- Execution time of MapReduce-MPI and different metadata on 1024 cores of Gordon (SDSC)
- Our method is not sensitive to metadata clustering



Three datasets from Docking@Home simulations

- 23, 21, and 12 ligands dock into HIV, trypsin, and p38alpha
- Search across **56** datasets of **100,000** conformations each
- Compare selection based on lower energy vs. our method

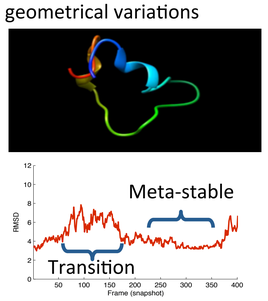
Our method shows significantly better accuracy in identifying native-like ligand conformations



Protein-folding Trajectories

Scientific Problem

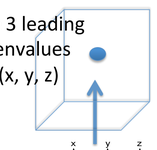
Enable clustering patterns in folding trajectories based on geometrical variations



Method

From protein conformations to metadata

- Use 3 leading eigenvalues for (x, y, z)



- Compute distance matrix using backbone 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

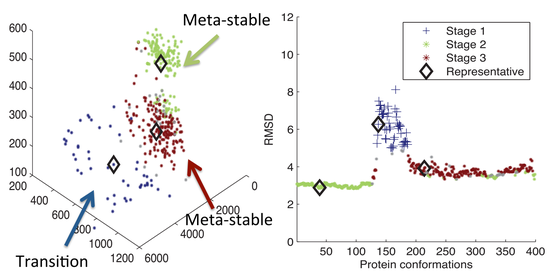
atoms

atoms

- Apply multi-dimensional scaling

	x	y	z
1	1.0	2.3	1.4
2	2.0	0.4	1.3
3	1.8	2.3	2.0
…	…	…	…
N	4.4	2.3	3.4

From metadata to scientific knowledge



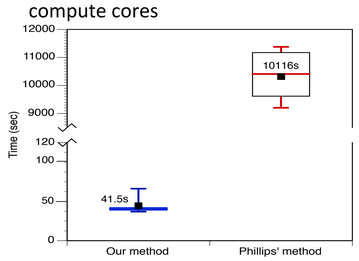
- Cluster set of points using hierarchical fuzzy c-means

- Compare with observed crystal structure

Results

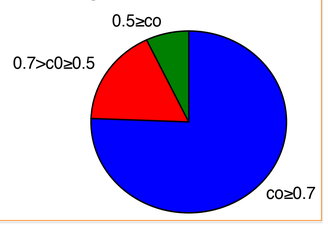
Compare our approach with traditional clustering method [Philip et al. 2013]

- Folding trajectories of villin headpiece subdomain (HP-35 NleNle)
- Parallel MATLAB on 256 Gordon compute cores



Study accuracy of 3D point mapping:

- 451 villin folding trajectories
- Assess linear correlation (co) between protein conformations and 3D points using Pearson test – strong correlation for co>0.5



Conclusion

We proposed a general method for classification and clustering analysis for large computational structural biology datasets on large distributed memory systems

Work in Progress

RNA Secondary Structures:

- Extend chunking to explore inversions that are far away from each other and have unlimited length (mega chunk)

Protein-folding Trajectories:

- Evaluate performance scalability for larger platforms (>1K cores) and trajectory datasets (>1TB data)
- Study protein structures that include both alpha helix and beta sheet

Acknowledgment CCF-SHF 1318445 UD

[1] B. Zhang et al. Enhancement of Accuracy and Efficiency for RNA Secondary Structure Prediction by Sequence Segmentation and MapReduce. BMC Structural Biology Journal, 13(Suppl 1):S3, 2013.
[2] B. Zhang et al. On Efficiently Capturing Scientific Properties in Distributed Big Data without Moving the Data - A Case Study in Distributed Structural Biology using MapReduce. In CSE'13, Sydney, Australia.
[3] B. Zhang et al. Enabling In-situ Data Analysis for Large Protein Folding Trajectory Datasets. In IPDPS'14, Phoenix, Arizona, USA.