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

Boyu Zhang (PhD Student, 5^{th} year) and Michela Taufer (Advisor)
Department of Computer and Information Science
University of Delaware
Newark, Delaware, 19716
Email: bzhang, taufer@udel.edu

I. Motivation and contribution

Today, petascale platforms perform large-scale simulations and generate massive amounts of data in a distributed fashion at unprecedented rates. This massive amount of data presents new challenges for the scientists analyzing the data's scientific meaning. Specifically in case of classification and clustering of the data, traditional analysis methods require the comparison of single records with each other in an iterative process, and therefore involve moving data across nodes of the system. When both the datasets and the number of nodes increase, classification and clustering methods can put an increasing pressure on the storage and the bandwidth of the system; thus the methods become inefficient and do not scale. In general, when analyzing scientific data, we focus on specific properties of the data. For example, in structural biology datasets, properties include the molecular geometry or the location of a molecule in a docking pocket. We claim that these properties can be captured across the dataset concurrently by analyzing each single data record independently. Deriving from this statement, in this research project, we propose a transformative data analysis method that comprises of two general steps. The first step extracts concise properties or features of each data record in parallel and represents them as metadata. The second step performs the analysis (i.e., classification or clustering) on the extracted properties. Since our method naturally fits in the MapReduce paradigm; we adapt it for different MapReduce frameworks (i.e., Hadoop, MapReduce-MPI, and DataMPI). We use the frameworks for three scientific datasets of RNA secondary structures, ligand conformations, and folding proteins.

The contributions of this dissertation are as follows:

- We introduce a transformative, general data analysis method together with effective algorithms supported by a MapReduce-style parallel programming model. The method avoids moving data to a centralized server, enables classification and clustering on large-scale data in a distrusted fashion, and assures both scalability and accuracy of the analysis.
- We apply the method to three representative and diverse computational structural biology datasets. The different datasets are: (1) large datasets of RNA sequences to identify sequence features of RNAs and classify the secondary structures as more or less likely to occur; (2) large datasets of docked ligand conformations to identify geometrical features of the ligand conformations and cluster the geometries in groups with high probabilities of well-docking into a protein pocket; and (3) large datasets of folding trajectories to identify recurrent patterns in protein-folding trajectories and cluster the folding trajectories into meta-stable and transition stages.

II. Classification of RNA secondary structures

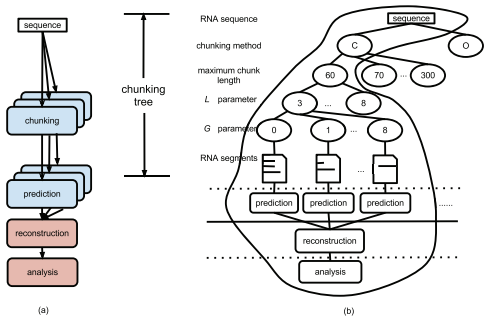


Figure 1: RNA secondary structure prediction workflow (a) an example of using statistical-based cutting method (b).

Given an RNA sequence, computational methods can predict multiple secondary structures, where each structure has a certain probability to happen in nature. The first analysis problem tackles in this research project is the classification of the secondary structures in more or less likely to occur based on the RNA sequence's family profile. To extract the secondary structures into metadata we cut the sequence into shorter chunks using statistical information, predict the secondary structure of each chunk independently using existing prediction programs, and reconstruct the whole secondary structures from the chunks' predictions. The workflow is

shown in Figure 1. The classification of secondary structures requires learning the RNA family profile from other secondary structures in the same family and classifying the new structures using this profile. The framework is implemented in Hadoop and evaluated using three datasets from the RFAM database and from the *Nodaviridae* virus dataset. The results show that our method exhibits linear scalability and can predict longer sequences than commonly used prediction programs. Moreover, our method generates more accurate secondary structures than the same prediction programs [1].

III. Clustering of Ligand Geometries

In drug design, when docking in a protein, ligands can more or less likely control protein's functions. The second analysis problem tackles the clustering of ligand geometries into sets with different probabilities of well-docking into the protein pocket. To extract the ligand's geometry into metadata we perform linear regression analysis on the ligand's atoms in which the coordinates are projected on three planes and interpolated (Figure 2). The clustering is an octree-based clustering by searching for the most dense subspace. It is expected that this subspace hosts well-docked ligands. The framework is implemented in MapReduce-MPI and evaluated using datasets generated by Docking@Home. The results show that our method achieves a linear scalability on 15 different datasets. Moreover, our method predicts more accurately well-docked ligand conformation than a more traditionally used hierarchical clustering method [2].

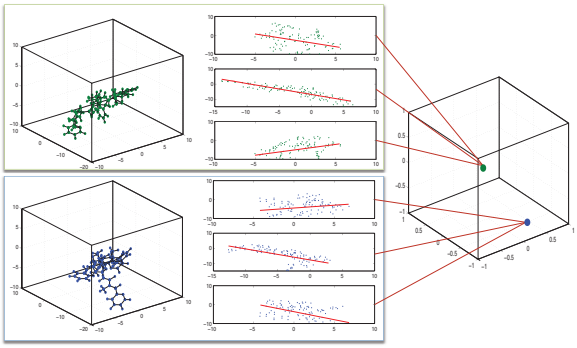
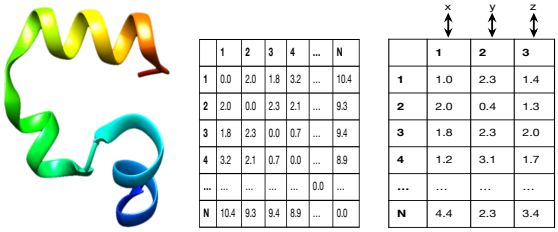


Figure 2: From ligand geometries to metadata.

IV. Clustering of Folding Trajectories

In protein folding, intra- and inter-trajectory analysis identify any folding patterns based on the geometric features of the folding conformations. The third analysis problem is to cluster these patterns and identify recurrent ones. We extract the geometric features into metadata by generating a distance matrix (DM) that represents the shape of each protein conformation, and performing a multi-dimensional scaling on the DM to generate a single 3D point as shown in



	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

	x	y	z
	1	2	3
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

(a) (b) (c)

Figure 3: One conformation of the villin HP-35 protein (a); part of its distance matrix using only its backbone atoms in the conformation (b); and three eigenvectors and the associated eigenvalues capturing and synthesizing the conformation geometry (c).

Figure 3. The clustering is performed as a hierarchical fuzzy c-means clustering on the set of 3D points to map a trajectory into meta-stable and transition stages. The framework is implemented in DataMPI, and evaluated using the folding trajectories datasets of protein HP-35 NleNle (i.e., a variant of the villin headpiece subdomain) and protein bovine pancreatic trypsin inhibitor (BPTI). The current results show that our method scales linearly when the datasets are up to 203GB. We test the accuracy on real datasets, while similar approaches normally worked on synthetic datasets [3].

V. Conclusion and Future Work

In this research project, we proved that it is possible to perform scalable classification and clustering analyses on large-scale datasets that are generated and stored in a distributed fashion. In addition, our method delivers more accurate results comparing to the traditional approaches. Moreover, our method can be applied to real world large datasets that are not studied before. Future work includes: applying the analysis method to a wider range of scientific datasets; exploring more methods to identify and extract relevant properties as well as to perform scalable classification and clustering analyses.

References

[1] B. Zhang, et al.. Enhancement of accuracy and efficiency for RNA secondary structure prediction by sequence segmentation and MR. *BMC Structural Biology*, 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 *Proc. of the IEEE CSE, 2013*.

[3] B. Zhang, et al.. Enabling in-situ data analysis for large protein-folding trajectory datasets. In Proc. of the *IEEE IPDPS, 2014*.