Data Analysis and Visualization Using HPC and Cloud

Judy Qiu, Zhenghao Gu, Bingjing Zhang, Yang Ruan, Thomas Wiggins

Indiana University

ABSTRACT

Our work aims to explore interoperability of large-scale cloud data processing software in HPC environments, and involves clustering and 3D visualization of gene sequence collections. We completed visualizing samples of 100K – 400K fungal sequences. This research involves major supercomputer computation as both clustering and visualization steps scale, even up to the square of the sample size. Our experiments were conducted on Indiana University's Big Red II [1] supercomputer and in collaboration with bioinformatics and biology faculty at IU.

BACKGROUND

In this experiment, the resultant clustering of fungal sequences profiles fungi species in a complex community. Phylogenetic analysis and clustering are two techniques which are commonly used to analyze sequence data. We want to compare the result from the pairwise clustering (visualized by Multi-Dimensional Scaling) with phylogenetic analysis. In our demo, each data point represents one fungus species.

The fungal sequences come from the variable D2 domain of the 28S Ribosomal RNA (rRNA) gene. All sequences were from species of arbuscular mycorrhizal fungi (AMF) because they exhibit a large amount of sequence variation, which can make them challenging to analyze. In this study the community structure of AMF in a clade of the genus Glomus was examined in undisturbed coastal grassland using large sub-unit (LSU) rDNA sequences amplified from roots of Hieracium pilosella.

MATERIALS AND METHODS

Harp [2] is a Hadoop plug-in made to abstract communication by transforming map-reduce programming models into map-collective models, reducing extraneous iterations and improving performance. Harp contains:

- Hierarchal data abstraction
- Collective communication model
- Pool-based memory management
- BSP-style Computation Parallelism
- Fault tolerance support with checkpointing

Our input data consists of sequences from the fungi ribosomal LSU. We use pairwise clustering and Multi-Dimensional Scaling (MDS) to perform large-scale sequence clustering and visualization. The results are compiled and displayed in PlotViz [3]. PlotViz is a 3D data point browser that displays large volumes of 2D or 3D data as points in a virtual space. Harp and PlotViz communicate via an ActiveMQ [4] broker, receiving and sending JAVA messages. This configuration decouples the data analyzing and visualization processes to guarantee smoothness of data plotting.

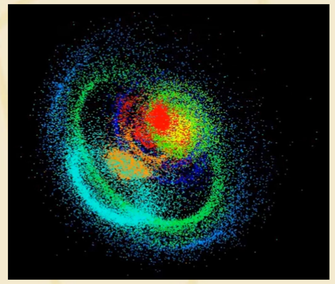
RESULTS TABLE

No. of Nodes	Data Size			
	100K (112 GB)	200K (448 GB)	300K (1006 GB)	400K (1.8 TB)
8	0.56 hr	NA	NA	NA
16	0.30 hr	NA	NA	NA
32	0.15 hr	0.74 hr	NA	NA
64	0.10 hr	0.44 hr	1.0 hr	NA
128	0.061 hr	0.22 hr	0.58 hr	0.92 hr

The experiments were carried out on Big Red II, which is a hybrid cluster with a total of 344 CPU nodes and 32 cores per node. We have developed a data clustering and visualization pipeline where the workflow deployed Hadoop to achieve maximum performance. The WDA-SMACOF algorithm runs with different problem sizes including 100K points, 200K, 300K and 400K. Because the input data is the distance matrix of points and related weight matrix and V matrix, the total size of input data is in quadratic growth: about 112 GB for a 100K problem, 448 GB for a 200K problem, 1 TB for 300K and 1.8 TB for 400K.

RESULTS

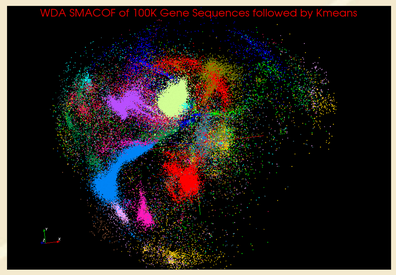
We cache distance matrix in short arrays, weight matrix in double arrays and V matrix in int arrays. Each point represents a gene sequence in a dataset of 454 pyrosequences from spores of known AMF species. The accompanying demo exhibits the affinities of fungi species by clustering. Sequences in the same cluster correspondingly show likelihood of genetic similarity in the phylogenetic tree as well.



RESULTS 2

The Scaling by Majorizing a Complicated Function (SMACOF) MDS algorithm is known to be fast and efficient. DA-SMACOF can reduce the time cost and find global optima by using deterministic annealing. The drawback is it assumes all weights are equal to one for all input distance matrices. To remedy this we added a weighting function to the SMACOF function, called WDA-SMACOF.

We used different MDS methods including WDA-SMACOF and EM-SMACOF. WDA-SMACOF uses Conjugate Gradient to avoid the cubic time complexity brought about by weighting and matrix inversion. The LMA usually had a result that was very similar to EM-SMACOF. Sequence differences were well preserved during MDS; WDA-SMACOF always had the lowest STRESS value compared to the other methods.



SUMMARY

This work explored a complex bioinformatics application associated with system architecture issues and challenges in order to develop a collective communication model supporting analysis of various Big Data problems.

We present a pipeline with pairwise clustering and MDS algorithms supported by Harp parallel data processing engine and PlotViz visualization tool which demonstrates the evolution of the data points and consequently exhibits the affinities of fungi species by clustering.

CONCLUSIONS

Our demonstration illustrates three novel/leading-edge technologies:

a) Clustering with deterministic annealing to obtain robust results avoiding local optima.
b) Dimension reduction using both deterministic annealing again (for robustness) and conjugate gradient to dramatically improve a matrix solver step (a factor of 5000 for a million sequences).
c) A novel cloud-HPC interoperability platform "Harp" that delivers high MPI quality parallel performance from a Hadoop platform.

We believe that both a) and b) are currently the best algorithms in their area and are implemented so they will give scalable data analytics on clouds and supercomputers.

REFERENCES

[1] Big Red II. Available at https://kb.iu.edu/d/bcqt

[2] Harp project. Available at http://salsaproj.indiana.edu/harp/index.html

[3] PlotViz. Available at http://salsahpc.indiana.edu/pviz3/

[4] ActiveMQ project, available at http://activemq.apache.org/