SCALABLE AND ROBUST DIMENSION REDUCTION AND CLUSTERING

Yang Ruan

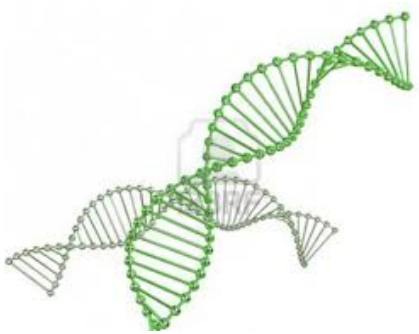
Advised by Geoffrey Fox

Outline

- Motivation
- Research Issues
- Experimental Analysis
- Conclusion and Futurework

Motivation

- Data Deluge
  - Increasing data size
  - Requires high performance computation power
- High Dimension Data
  - Verify clustering result



.....

Motivation (2)

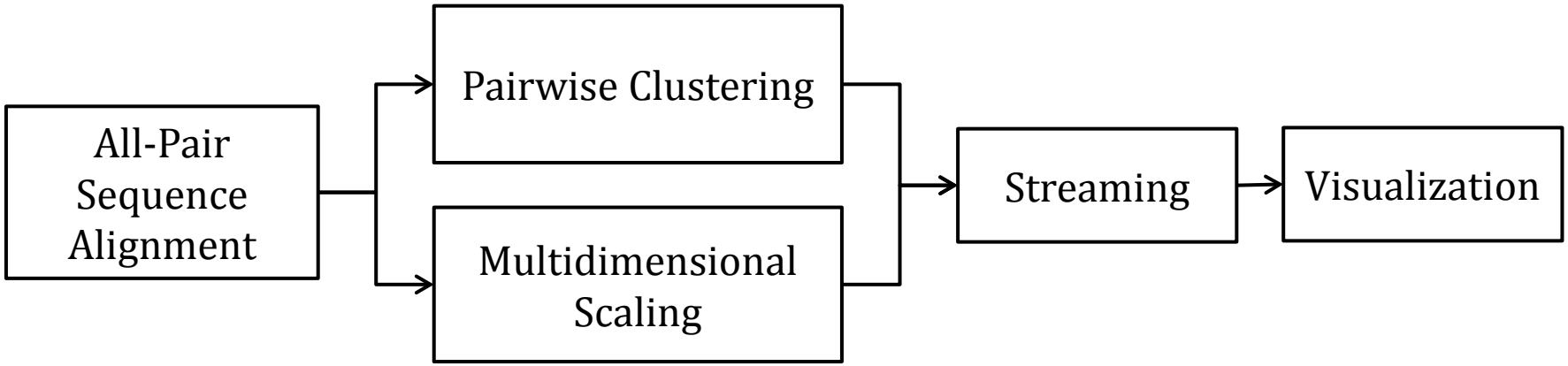
- Multidimensional Scaling
  - Cluster large-scale of different kinds of data
  - Visualize the result in 3D
  - Traditional $O(N^2)$ methods doesn't work
- Interpolation/ Streaming
  - O(N) method
  - Can be pleasingly paralleled
  - Lower precision but faster speed
- Phylogenetic Tree Visualization
  - Slow when data size increases
  - Traditional display method doesn't work with clustering

Outline

- Motivation
- Research Issues
  - Overview of Workflow
  - DA-SMACOF with Weighting
  - Hierarchical Interpolation with Weighting
  - 3D Phylogenetic Tree Display with Clustering
- Experimental Analysis
- Conclusion and Futurework

Overview of Workflow

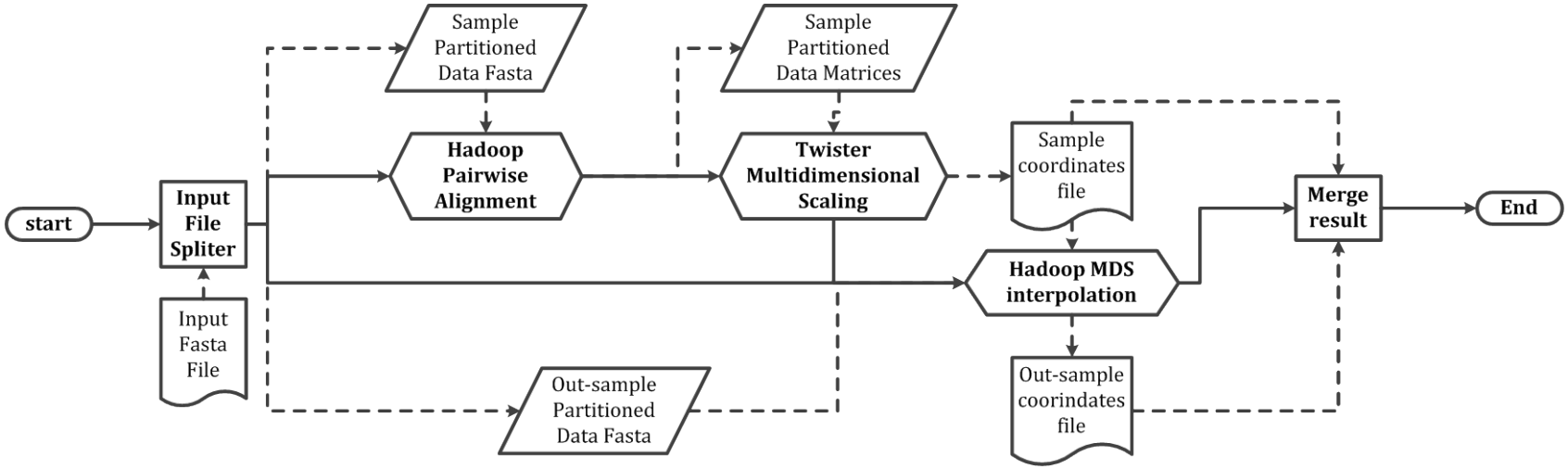
- Deterministic Annealing Clustering and Interpolative Dimension Reduction Method (DACIDR)
  - Dimension Reduction (Multidimensional Scaling)
  - Clustering (Pairwise Clustering)
  - Streaming/ Interpolation



Simplified Flow Chart of DACIDR

Workflow Parallelization

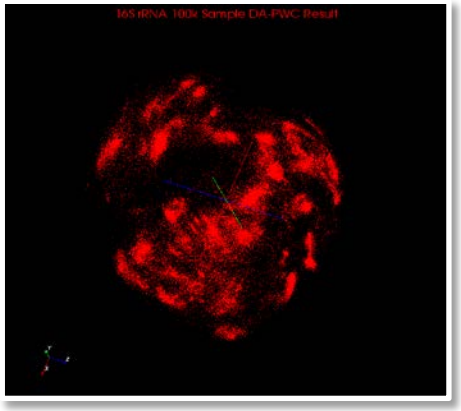
- Hybrid MapReduce workflow (HyMR)
  - Use Hadoop for all-pair sequence alignment and interpolation
    - Faster execution using dynamic scheduling for finer granularity tasks
    - Fault tolerance while executing long running jobs.
  - Use Twister for multidimensional scaling
    - Faster execution for iterative algorithms



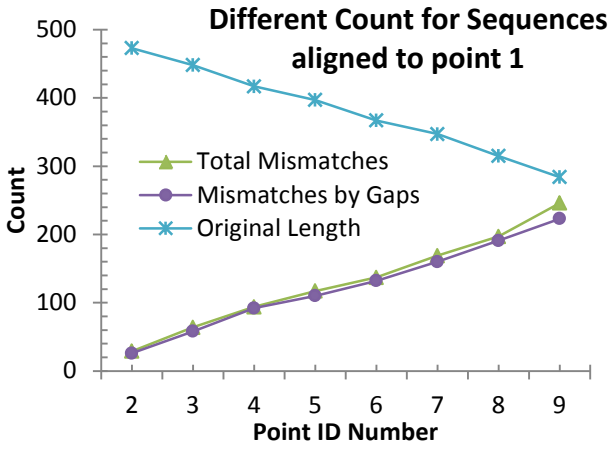
Flow Chart of DACIDR running on HyMR

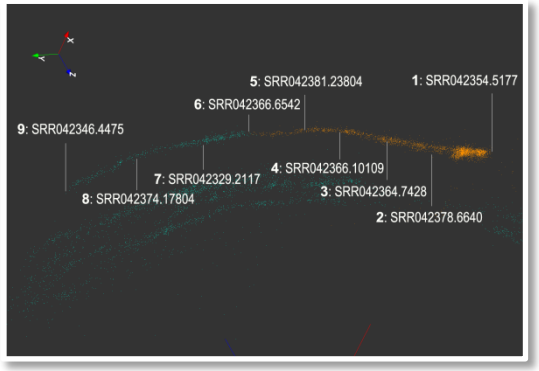
Distance Calculation

- Smith Waterman performs local alignment while Needleman Wunsch performs global alignment.
- Global alignment suffers problem from various lengths in this dataset

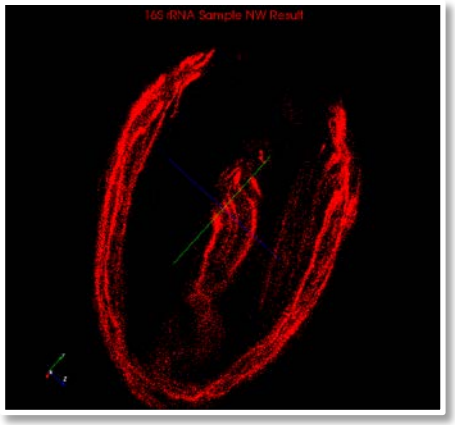


Visualization result using SW





One "Cigar" selected from NW visualization result



Visualization result using NW

Possible Issues

- Local sequence alignment (Smith Waterman) could generate very low quality distances.
  - E.g. For two sequences with original length 500, it could generate an alignment with length 10 and gives a pid of 1 (distance 0) even if these two sequences shouldn't be near each other.
- Sequence alignment is time consuming.
  - E.g. To interpolate 100k *out-of-sample* sequences (average length of 500) into 10k in-sample sequences took around 100 seconds to finish on 400 cores, but to align them took around 6000 seconds to finish on same number of cores.

Outline

- Motivation
- Research Issues
  - Overview of Workflow
  - DA-SMACOF with Weighting
  - Hierarchical Interpolation with Weighting
  - 3D Phylogenetic Tree Display with Clustering
- Experimental Analysis
- Conclusion and Futurework

WDA-SMACOF (background)

- Multidimensional Scaling
  - Given proximity data in high dimension space.
  - Non-linear optimizing problem to find mapping in target dimension space by minimizing an object function.
  - Object function is often given as STRESS or SSTRESS:

$$\sigma(\boldsymbol{X}) = \sum_{i<j\leq N} w_{ij}(d_{ij}(\boldsymbol{X}) - \delta_{ij})^2 \qquad (1)$$

$$\sigma^2(\boldsymbol{X}) = \sum_{i<j\leq N} w_{ij}[(d_{ij}(\boldsymbol{X}))^2 - (\delta_{ij})^2]^2 \qquad (2)$$

where X is the mapping in the target dimension, $d_{ij}(X)$ is the dissimilarity between point and point in original dimension space, w_{ij} denotes the possible weight from each pair of points that , δ_{ij} denotes the Euclidean distance between point and in target dimension.

WDA-SMACOF (background 2)

- Scaling by Majorizing a Complicated Function (SMACOF)
  - An EM-like algorithm that decreases STRESS iteratively
  - Could be trapped in local optima
- DA-SMACOF
  - Use Deterministic Annealing to avoid local optima
  - Introduce a computational temperature T.
  - By lowering the temperature during the annealing process, the problem space gradually reveals to the original object function.
  - Assume all weights equals 1.
- Conjugate Gradient
  - An iterative algorithm that solves linear equations.
  - CG is used to solve Ax=b where x and b are both vectors of length N and A is an N * N symmetric positive definite (SPD) matrix.

WDA-SMACOF

- When distance is not reliable or missing, set the weight correspond to that distance to 0.
- Similar to DA-SMACOF, the updated STRESS function is derived as

$$\sigma(X_T) = \sum_{i<j\leq N} w_{ij}\,(d_{ij}(X_T) - \tilde{\delta}_{ij})^2 \qquad (3)$$

where $\tilde{\delta}_{ij}$ is defined as

$$\tilde{\delta}_{ij} = \begin{cases} \delta_{ij} & if\ w_{ij} = 0 \\ \delta_{ij} - \sqrt{2TL} & else\ if\ \delta_{ij} > \sqrt{2TL} \\ 0 & other\ wise \end{cases} \qquad (4)$$

- When T is smaller, $\tilde{\delta}_{ij}$ is larger, so the original problem space is gradually revealed.

WDA-SMACOF (2)

- By deriving a majorizing function out of the STRESS function, the final formula is:

$$VX_T = B(Z_T)Z_T \tag{5}$$

$$X_T^u = V^{\dagger}B(Z_T)Z_T \tag{6}$$

where V and $B(X_T)$ is defined as following:

$$v_{ij} = \begin{cases} -w_{ij} & if\ i \neq j \\ -\sum_{k \neq i} v_{ik} & else\ if\ i = j \end{cases} \tag{7}$$

$$b_{ij} = \begin{cases} -\frac{w_{ij}\tilde{\delta}_{ij}}{d_{ij}(X_T)} & if\ i \neq j, w_{ij} \neq 0, and\ d_{ij}(X_T) \neq 0 \\ -\sum_{k \neq i} b_{ik} & else\ if\ i = j \\ 0 & otherwise \end{cases} \tag{8}$$

$V^{\dagger}$ is the pseudo-inverse of V.

WDA-SMACOF (3)

- Pseudo-Inverse of V is given as $(V+11')^{-1} - n^{-2}11'$
  - Matrix Inversion has a time complexity of $O(N^3)$
    - Cholesky Decomposition, Singular Vector Decomposition…
    - Traditional SMACOF matrix inversion is trivial for small dataset
- Use Conjugate Gradient (CG) to solve $VX=B(Z)Z$
  - X and B(Z)Z are both N * L matrix and V is N * N matrix.
  - denote $V + I$ as $\dot{V}$

Theorem 1. $\dot{V}$ is a symmetric positive definite (SPD) matrix.

Proof. Since $w_{ij} = w_{ji}$, so $v_{ij} = v_{ji}$, and $\dot{V} = \dot{V}^T$. From (17), $\dot{V}$ can be represented as

$$\dot{v}_{ij} = \begin{cases} -w_{ij} & if\ i \neq j \\ 1 + \sum_{k \neq i} w_{ik} & else\ if\ i = j \end{cases} \quad (24)$$

Because $w_{ij} \geq 0$, so $\dot{v}_{ii} > 0$. And $\dot{v}_{ii} > \sum_{k \neq i} w_{ik} = \sum_{k \neq i} |\dot{v}_{ik}|$. So according to [24], Theorem 1 is proved.

WDA-SMACOF (4)

- Conjugate Gradient
  - Denote r_i as the residual and d_i as the direction in *ith* iteration
  - X can be updated using
  - Only a number of iterations << N will are needed for approximation

$$\alpha_i = \frac{dot(r_i^t, r_i)}{dot(d_i^t, \dot{V} d_i)} \quad (9)$$

$$X_{i+1} = X_i + \alpha_i d_i \quad (10)$$

$$r_{i+1} = r_i - \alpha_i \dot{V} d_i \quad (11)$$

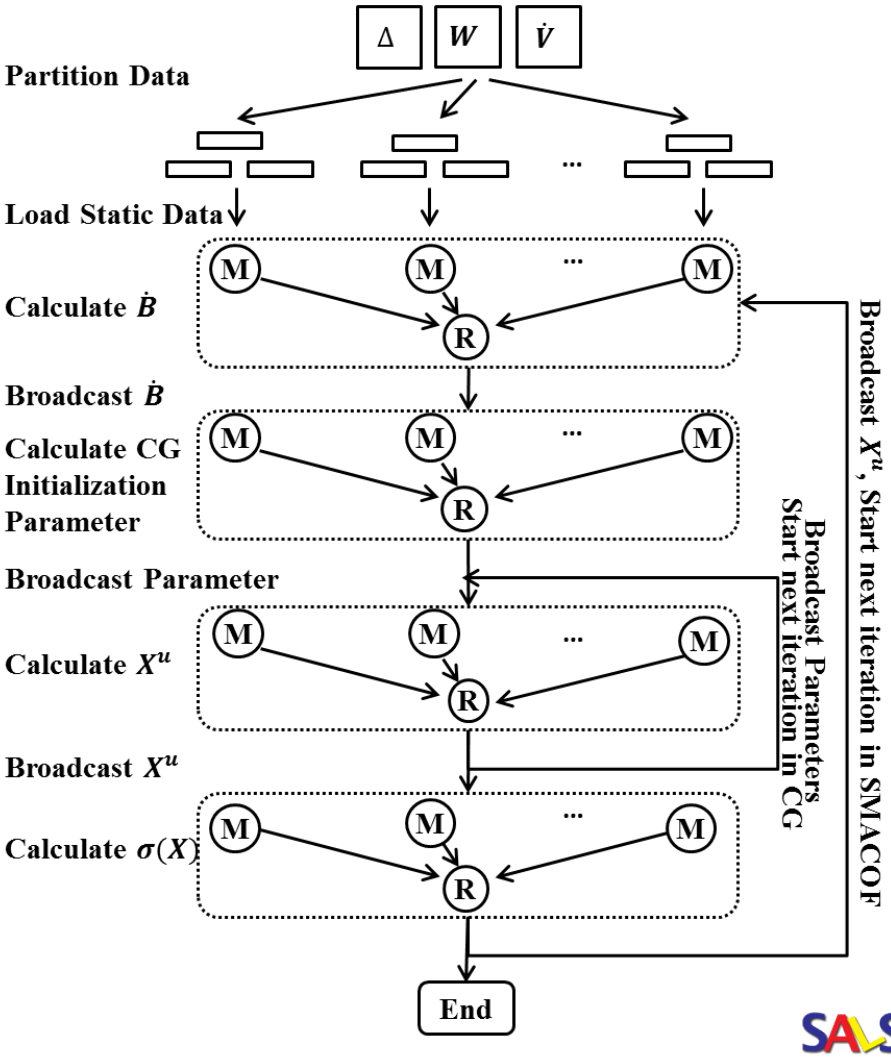
$$\beta_{i+1} = \frac{dot(r_{i+1}^t, r_{i+1})}{dot(r_i^t, r_i)} \quad (12)$$

$$d_{i+1} = r_{i+1} + \beta_{i+1} d_i \quad (13)$$

where $dot(X, Y) = \sum_{1 \le j \le L} \sum_{1 \le i \le N} x_{ji} y_{ij}$

WDA-SMACOF (5)

- Parallelization
  - Parallelized using Twister, an iterative MapReduce runtime
  - One outer loop is for one SMACOF iteration
  - One inner loop is for one CG iteration
  - Time complexity $O(N * N * l_1 * l_2)$



Outline

- Motivation
- Research Issues
  - Overview of Workflow
  - DA-SMACOF with Weighting
  - Hierarchical Interpolation with Weighting
  - 3D Phylogenetic Tree Display with Clustering
- Experimental Analysis
- Conclusion and Futurework

Interpolation (background)

- Out-of-sample / Streaming / Interpolation problem
  - Original MDS algorithm needs $O(N^2)$ memory
  - Given an *in-sample* data result, interpolate *out-of-sample* into the *in-sample* target dimension space.
- Majorizing Interpolation MDS (MI-MDS)
  - Based on pre-mapped MDS result of n sample data.
  - Find a new mapping of the new point based on the position of k nearest neighbors (k-NN) among n sample data.
  - Iterative majorization method is used.
  - Needed O(MN) distance computation
  - Assume all weights equal one.

Interpolation

- MI-MDS with weighting and deterministic annealing (WDA-MI-MDS)
  - Adding weight to object function, where each weight correspond to a distance from an *out-of-sample* point to an *in-sample* point.
  - Update STRESS function:

$$\sigma(X) = \sum_{i \le N} w_{i\hat{x}} \left(d_{i\hat{x}}(X) - \delta_{i\hat{x}}\right)^2 \qquad (14)$$

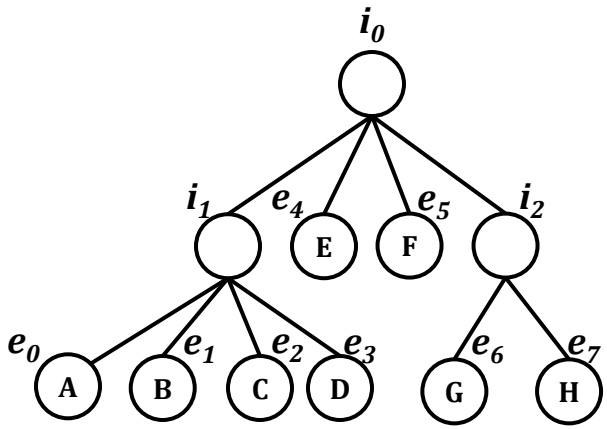
  - Adding a computational temperature T as same as in DA-SMACOF
  - Final formula is updated as

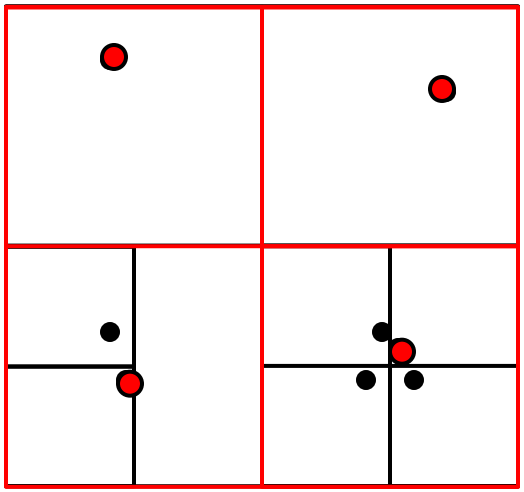
$$\hat{x}_t^u = \frac{q^t + \sum_{i \le N} \frac{w_{i\hat{x}} \tilde{\delta}_{i\hat{x}_t}}{d_{iz}} (z - p_i)}{\sum_{i \le N} w_{i\hat{x}}} \qquad (15)$$

where $q^t = \left(\sum_{i \le N} w_{ix} p_{i1}, \ldots \sum_{i \le N} w_{ix} p_{iL}\right)$

Interpolation (2)

- Hierarchical Interpolation
  - Sample Sequence Partition Tree
    - SSP-Tree is an octree to partition the *in-sample* sequence space in 3D.
  - Closest Neighbor Tree
    - CN-Tree is a hyper space tree that partition the *in-sample* sequences by their original distances.

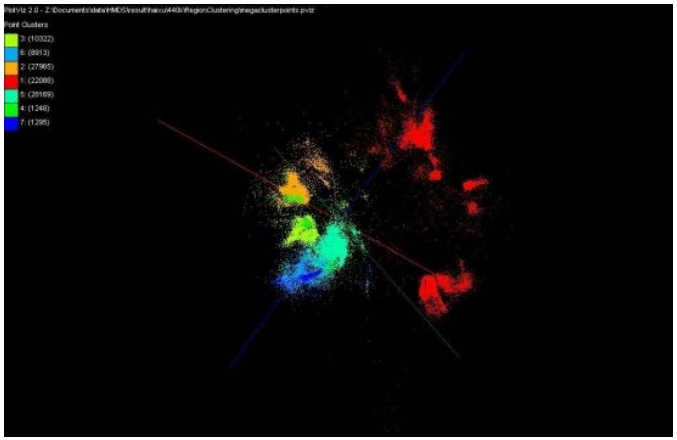




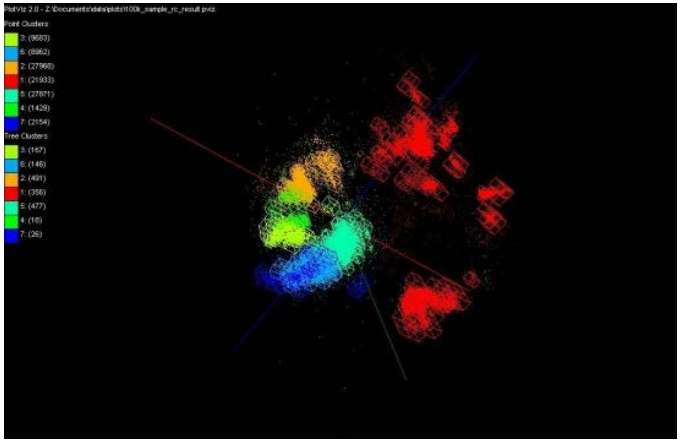
An example for SSP-Tree in 2D with 8 points

Interpolation (3)

- HI-MI compare with each center point of each tree node, and searches k-NN points from top to bottom.
  - Error between original distance and target dimension distance
- HE-MI use more sophistic heuristic
  - Use a heuristic function to determine a quality of a tree node.
  - May increase search levels in tree to reduce time cost.
  - Computation complexity
  - MI-MDS: O(NM)
  - HI-MI: O(MlogN)
  - HE-MI: $O(M(N^T + T))$



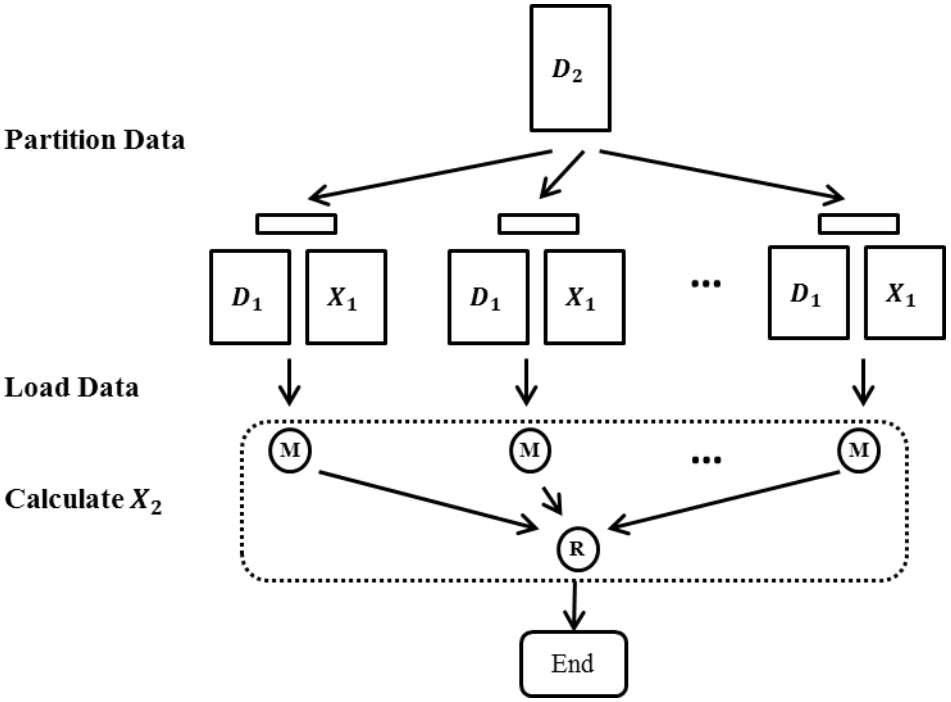
100k data after dimension reduction in 3D



100k data with tree nodes displayed

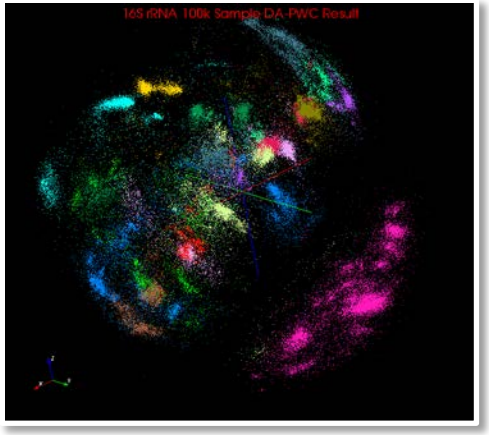
Interpolation (4)

- Parallelization
  - Pleasingly parallel application
  - Can be parallelized by either Twister or Hadoop

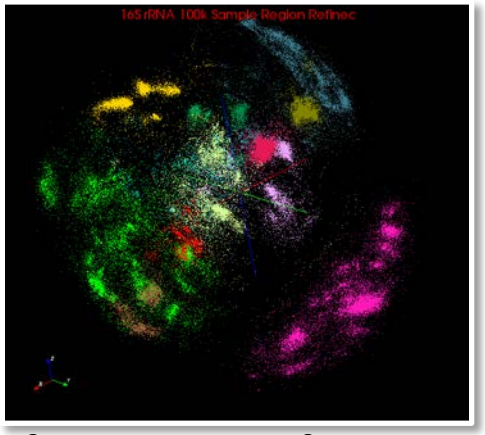


Region Refinement

- An algorithm uses Oct-tree to refine the clustering results in 3D.
  - Similar to k-means, update the centers after points in clusters are re-assigned
  - Only needs to compare tree node center instead of every points to centroid.



Before Region Refinement



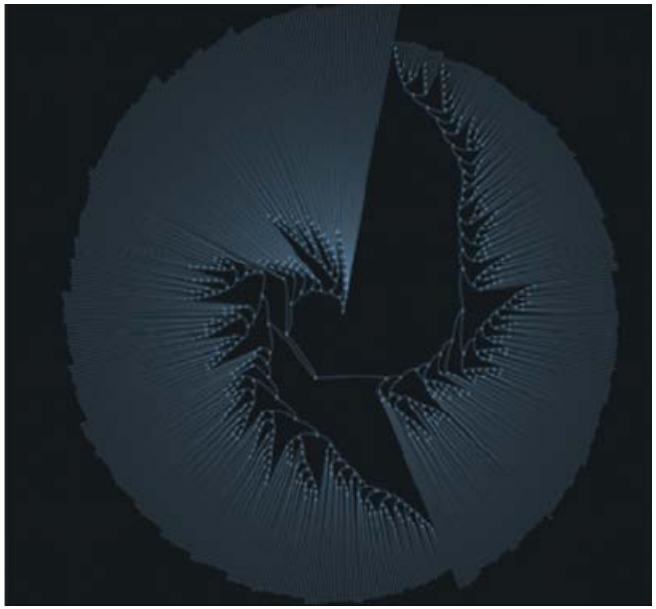
After Region Refinement

Outline

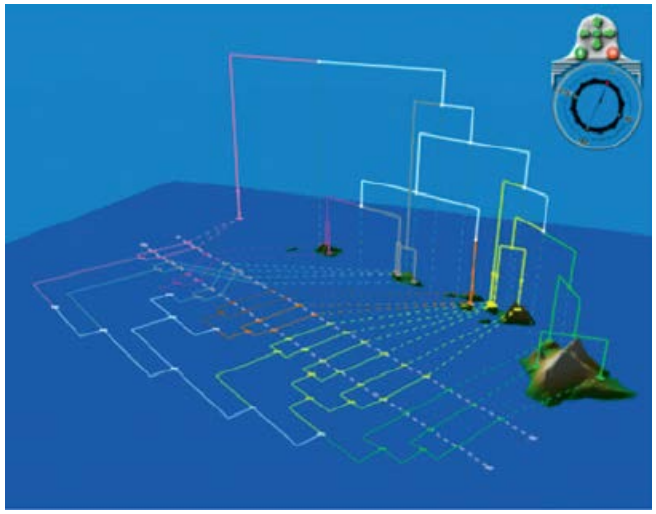
- Motivation
- Research Issues
  - Overview of Workflow
  - DA-SMACOF with Weighting
  - Hierarchical Interpolation with Weighting
  - 3D Phylogenetic Tree Display with Clustering
- Experimental Analysis
- Conclusion and Futurework

Phylogenetic Tree Visualization (Background)

- Phylogenetic Tree Display
  - 2D/3D display, such as rectangular cladogram, circular phylogram.
  - Only preserves the proximity of children and their parent.



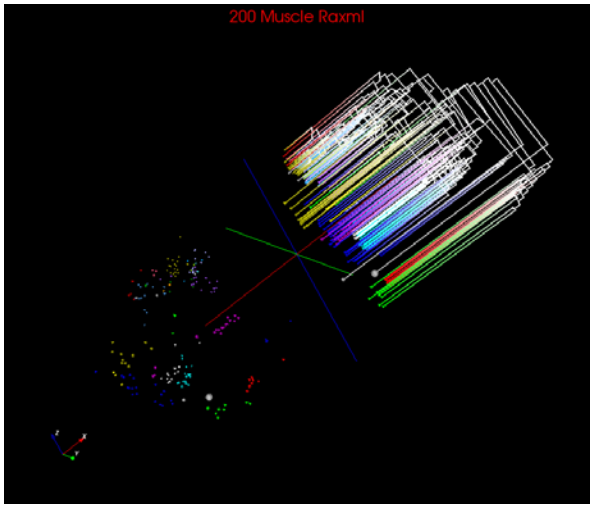
Circular Phylogram

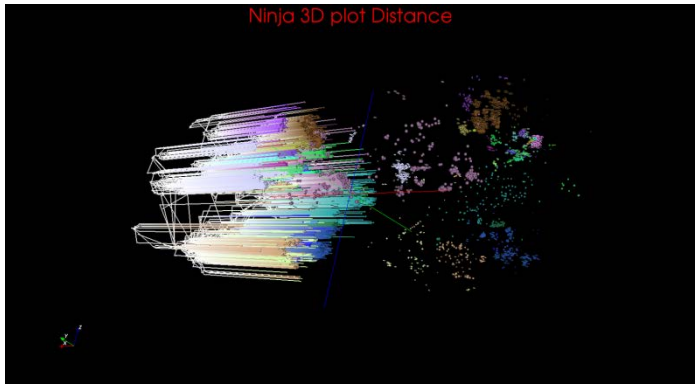


Rectangular Cladogram

3D Phylogenetic Tree Visualization

- Phylogenetic Tree Generation
  - Generate a phylogenetic tree, e.g. Multiple Sequence Alignment and RaXml, Pairwise Sequence Alignment and Ninja
- Cubic Cladogram
  - Use Principle Component Analysis (PCA) to select a plane which has the largest eigenvalue.
  - For each point in the 3D space, project a point onto that plane
  - Generate internal nodes of the tree by projecting them onto the edges from tree top to bottom.





Cuboid Cladogram Examples

3D Phylogenetic Tree (2)

- Spherical Phylogram
  - Select a pair of existing nodes *a* and *b*, and find a new node *c*, all other existing nodes are denoted as *k*, and there are a total of *r* existing nodes. New node *c* has distance:

$$d(a,c) = 0.5 * d(a,b) + \frac{\sum_{k=1}^{r} [d(a,k) - d(b,k)]}{2(r-2)} \quad (16)$$

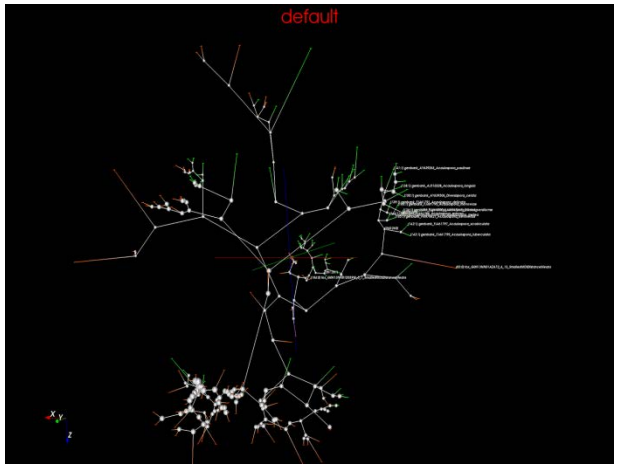
$$d(b,c) = d(a,b) - d(a,c) \quad (17)$$

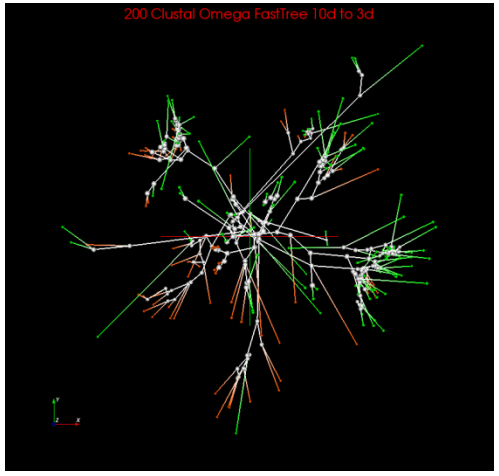
$$d(c,k) = 0.5 * [d(a,k) + d(b,k) - d(a,b)] \quad (18)$$

  - The existing nodes are *in-sample* points in 3D, and the new node is an *out-of-sample* point, thus can be interpolated into 3D space.

3D Phylogenetic Tree (3)

- Spherical Phylogram
  - The tree is generated from bottom to top
  - Various distance measure
    - 3D distance
    - Original distance
    - Two step dimension reduction distance, i.e. original distance to 10D, 10D to 3D.
  - Various tree
    - Existing tree, e.g. From RaXml
    - Generate tree, i.e. neighbor joining
  - Finds global optima





Spherical Phylogram Examples

Outline

- Motivation
- Background and Related Work
- Research Issues
- Experimental Analysis
  - WDA-SMACOF and WDA-MI-MDS
  - Hierarchical Interpolation
  - 3D Phylogenetic Tree Display
- Conclusion and Futurework

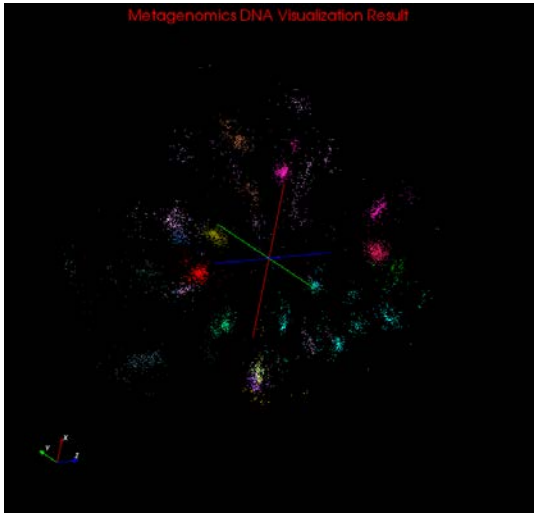
Experimental Environment

- Environment
  - 100 nodes (800 cores) of PolarGrid
  - 80 nodes (640 cores) of FutureGrid Xray
  - 128 nodes (4096 cores) of BigRed2
- Dataset
  - 16S rRNA data with 1.1 million sequences
  - Metagenomics data with 4640 sequences
  - COG Protein data with 183k sequences
- Parallel Runtimes
  - Hadoop, an open source MapReduce runtime
  - Twister, an iterative MapReduce runtime

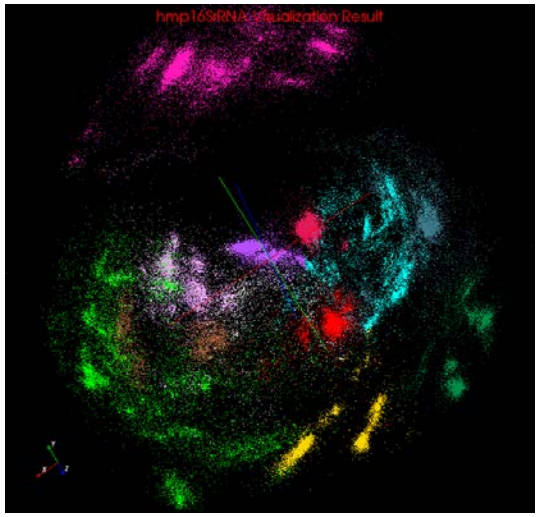
Visualization

- Use PlotViz3 to visualize the result
- Different colors are from clustering result

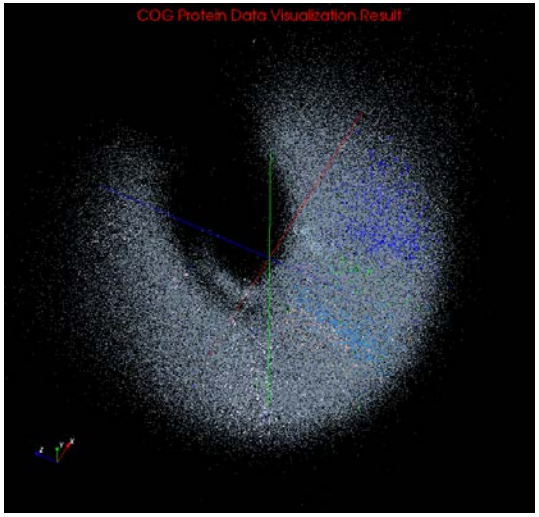
Metagenomics



hmp16SrRNA

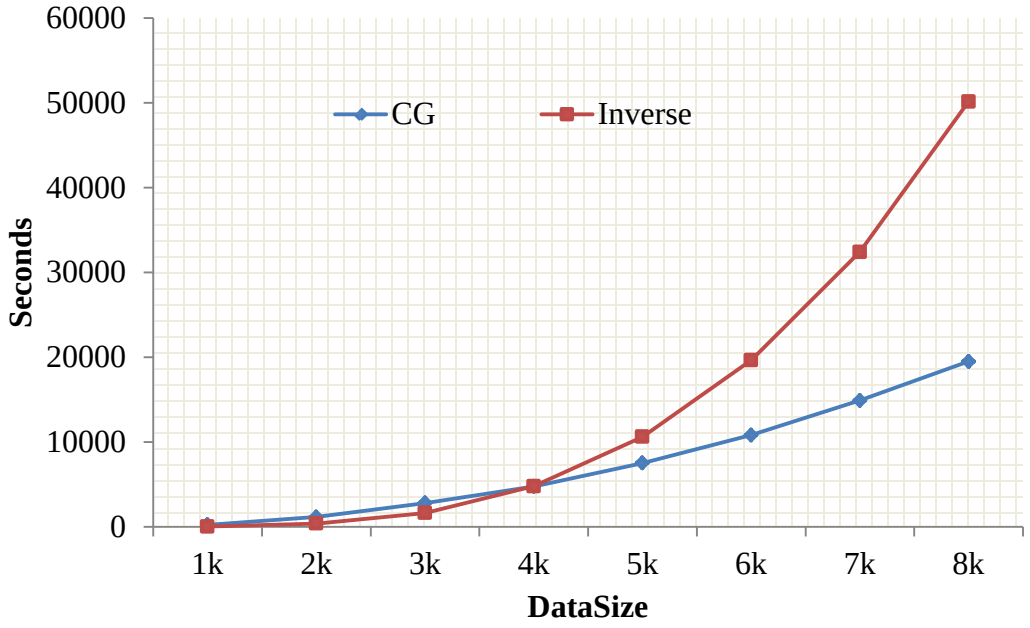


COG Protein



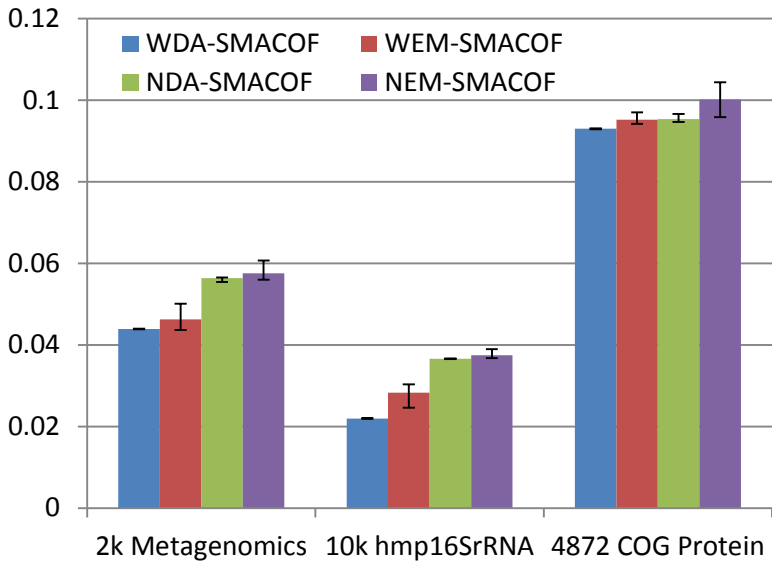
WDA-SMACOF

- Time Cost of CG vs Matrix Inversion (Cholesky Decomposition)
  - Input Data: 1k to 8k hmp16SrRNA sequences
  - Environment: Single Thread



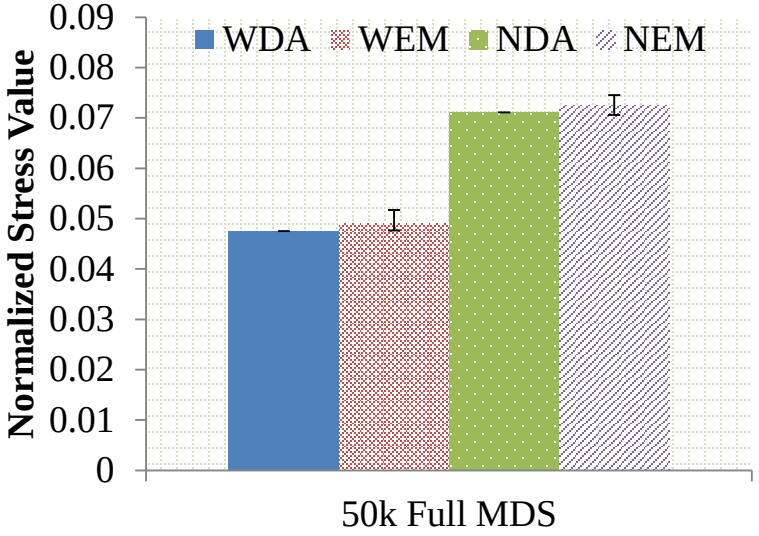
WDA-SMACOF (2)

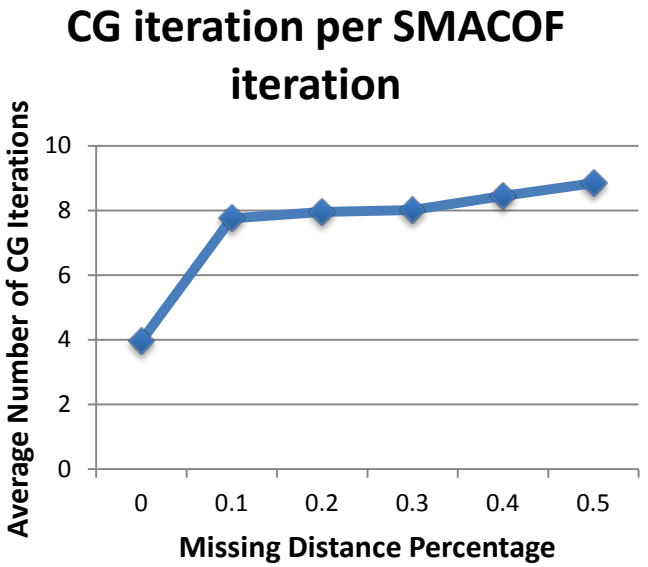
- Normalized STRESS of WDA-SMACOF vs DA-SMACOF and EM-SMACOF
  - Input Data: 2k Metagenomics DNA, 10k hmp16SrRNA, and 4872 COG Protein sequences.
  - Running Environment: FutureGrid Xray from 80 cores to 320 cores.
  - 10% of distances are considered missing



WDA-SMACOF (3)

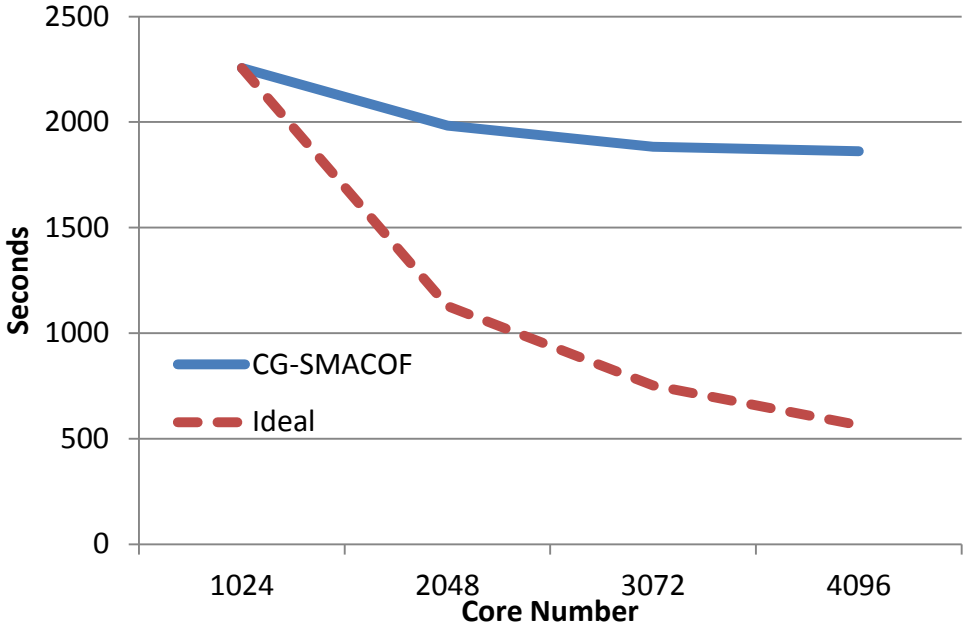
- Large Scale Test for WDA-SMACOF
  - Input Data: 50k hmp 16S rRNA sequences
  - Environment: FutureGrid Xray with 640 cores.





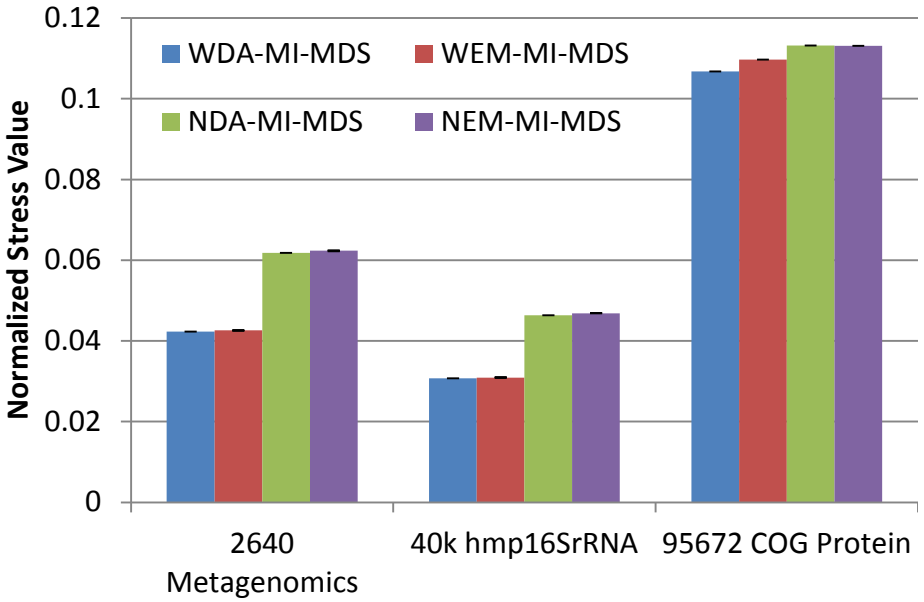
WDA-SMACOF (4)

- Time Cost of Strong scale up for WDA-SMACOF
  - Input Data: 100k hmp16SrRNA sequences
  - Environment: 32 nodes (1024 cores) to 128 nodes (4096 cores) on BigRed2.
    - Will complete this graph with 256 cores, 512 cores and 768 cores.



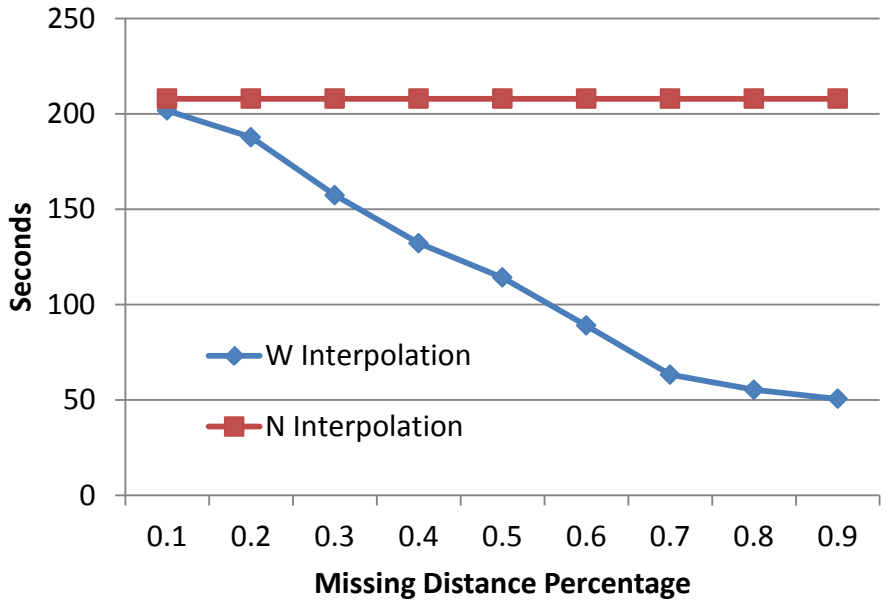
Interpolation (1)

- Normalized STRESS value of WDA-MI-MDS vs MI-MDS and other methods
  - Input Data: 2640 out-of-sample Metagenomics Sequences to 2k in-sample sequences, 40k out-of-sample hmp16SrRNA sequences to 10k in-sample sequences, 95672 out-of-sample COG Protein sequences to 4872 in-sample sequences.
  - Environment: FutureGrid Xray, 80 cores to 320 cores.
  - 10% of distances are missing



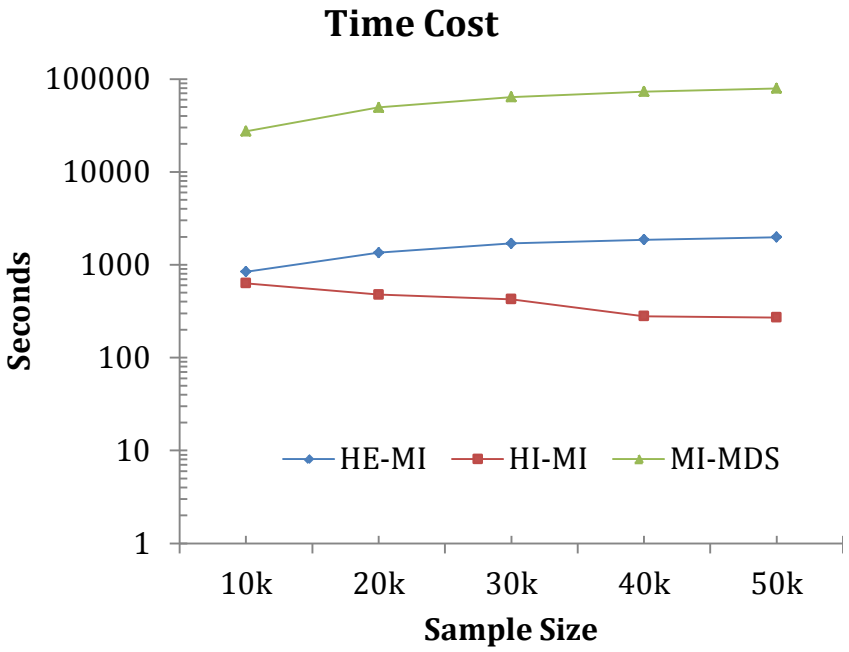
Interpolation (2)

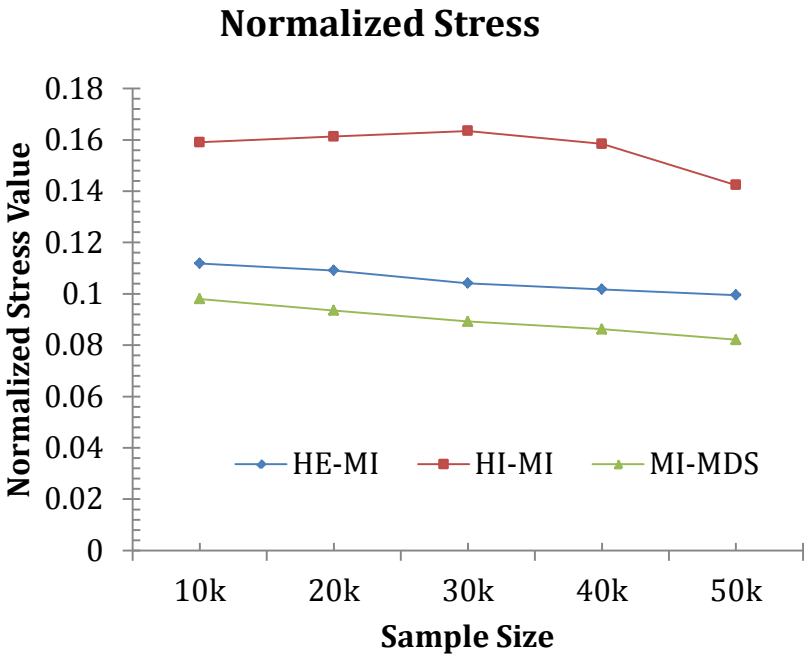
- Time cost of Weighted Interpolation vs Non-weighted Interpolation
  - Input Data: Interpolate 40k out-of-sample into 10k in-sample hmp16SrRNA sequences.
  - Increasing missing distance from 10% to 90%
  - Fixed to 400 iterations



Interpolation (3)

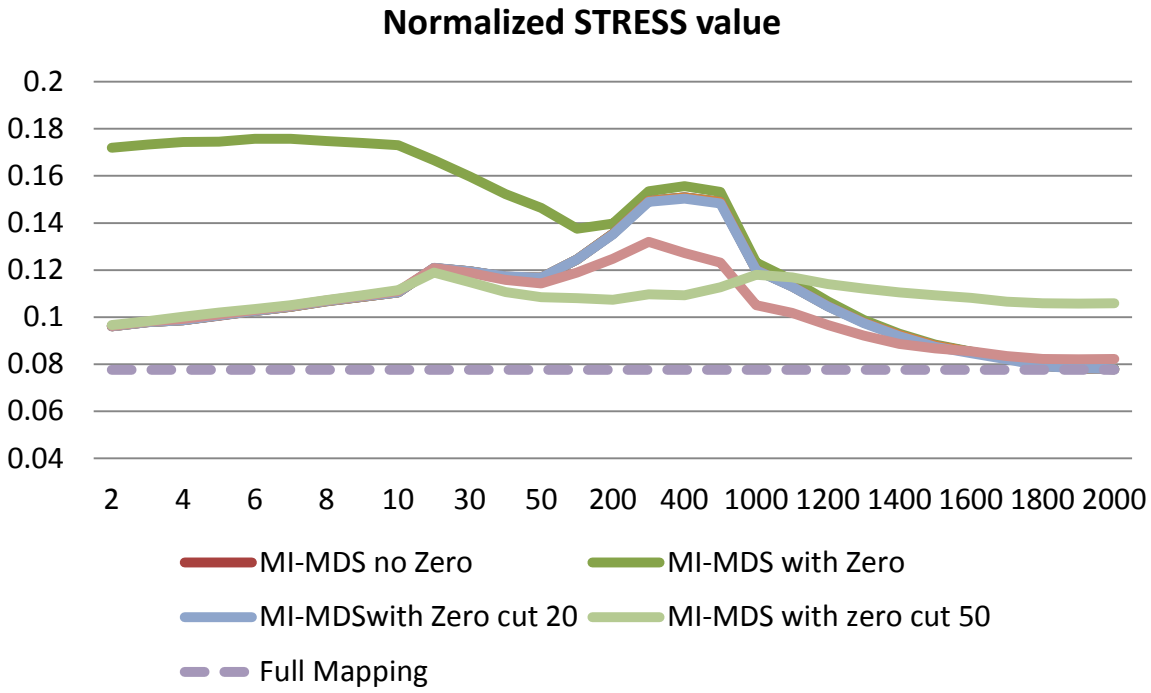
- Normalized STRESS and time cost of HE-MI vs HI-MI and MI-MDS
  - Input set: 100k hmp16SrRNA sequences
  - Environment: 32 nodes (256 cores) from PolarGrid





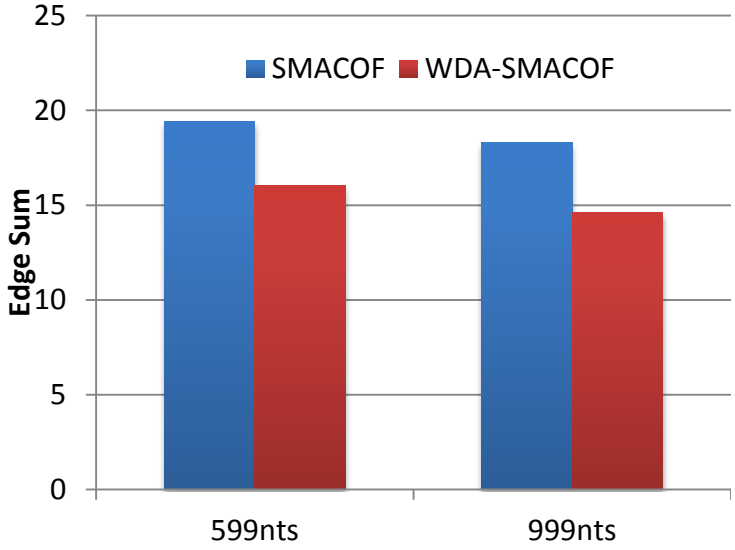
Interpolation (4)

- Normalized STRESS value by increasing k-NN for MI-MDS
  - Input Data: Interpolated 2640 out-of-sample Metagenomics DNA data to 2k in-sample data
  - Environment: 4 nodes (32 cores) of PG

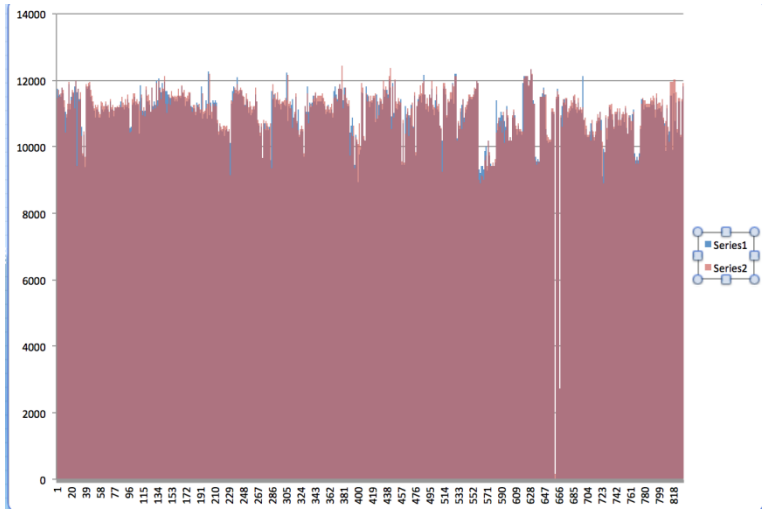


Phylogenetic Tree

- Spherical Phylogram using different dimension reduction methods
  - Edge Sum
    - Sum over all the length of edges
  - Local Optima (examples)
    - FR750020_Arc_Sch_K
    - FR750022_Arc_Sch_K



Original distances from FR750020_Arc_Sch_K and FR750022_Arc_Sch_K to all other 832 points.



Outline

- Motivation
- Background and Related Work
- Research Issues
- Experimental Analysis
- Conclusion and Futurework

Conclusion

- Distance measurement is essential.
- WDA-SMACOF can has higher precision with sparse data much better than DA-SMACOF with time complexity of $O(N^2)$.
- WDA-MI-MDS has higher precision with sparse data than MI-MDS.
- HE-MI has a slight higher stress value than MI-MDS, but much lower time cost, which makes it suitable for massive scale dataset.
- 3D phylogenetic tree with clustering enables easy observation of data.

Futurework

- Enable WDA-SMACOF to process more than 100k sequences
- Test WDA-SMACOF using Sammon's STRESS
- Improve accuracy of HE-MI
- Improvement over 3D phylogenetic trees.
- Release a stable WDA-SMACOF and WDA-HE-MI workflow which can be directly submitted through TORQUE
- Release 3D phylogenetic tree spherical phylogram and cubic cladogram generation software.

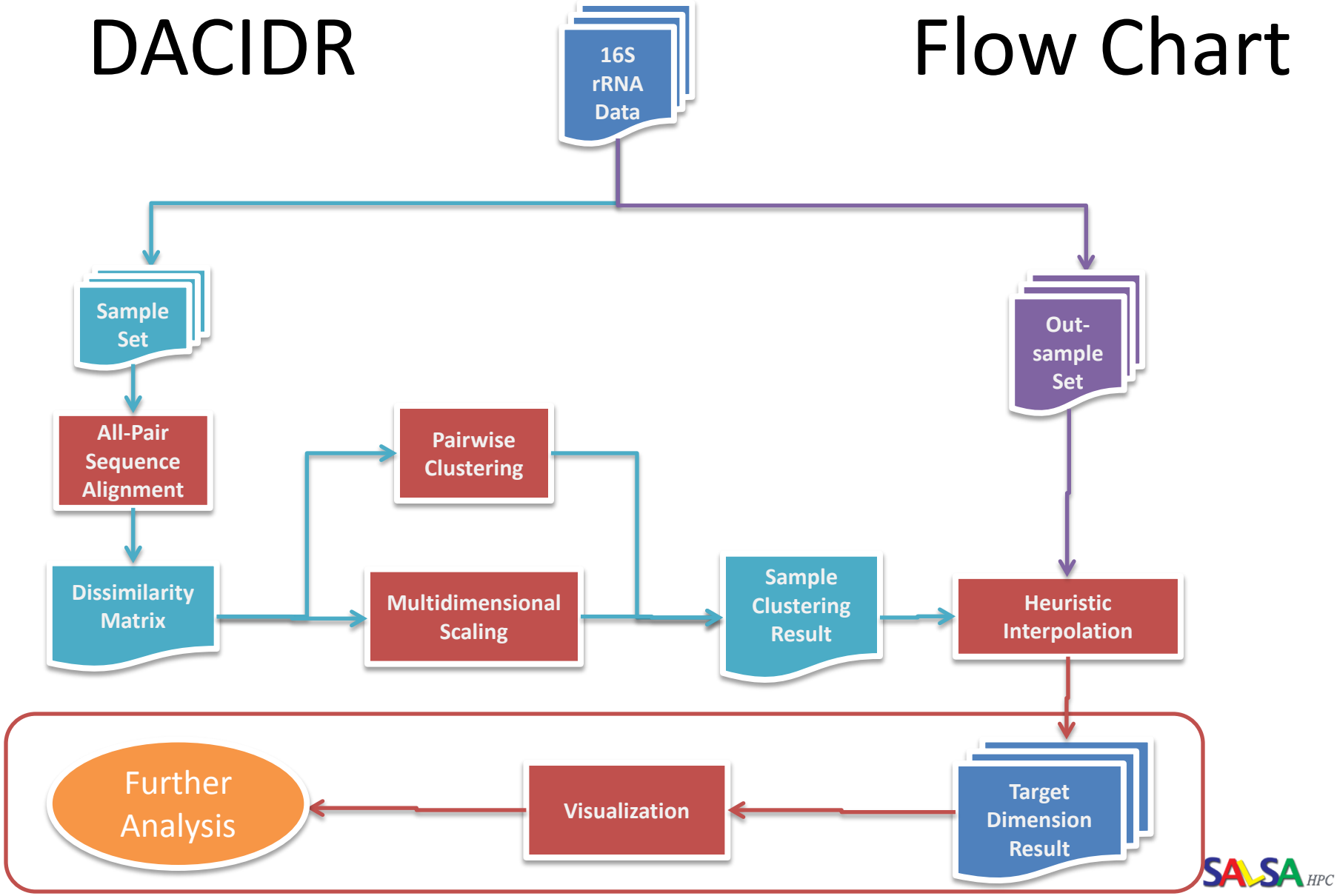
Reference

- Yang Ruan, Geoffrey Fox. **A Robust and Scalable Solution for Interpolative Multidimensional Scaling with Weighting**. Proceedings of IEEE eScience 2013, Beijing, China, Oct. 22-Oct. 25, 2013. (Best Student Innovation Award)
- Yang Ruan, Saliya Ekanayake, et al. **DACIDR: Deterministic Annealed Clustering with Interpolative Dimension Reduction using a Large Collection of 16S rRNA Sequences**. Proceedings of ACM-BCB 2012, Orlando, Florida, ACM, Oct. 7-Oct. 10, 2012.
- Yang Ruan, Zhenhua Guo, et al. **HyMR: a Hybrid MapReduce Workflow System**. Proceedings of ECMLS'12 of ACM HPDC 2012, Delft, Netherlands, ACM, Jun. 18-Jun. 22, 2012.
- Adam Hughes, Yang Ruan, et al. **Interpolative multidimensional scaling techniques for the identification of clusters in very large sequence sets**, BMC Bioinformatics 2012, 13(Suppl 2):S9.
- Jong Youl Choi, Seung-Hee Bae, et al. **High Performance Dimension Reduction and Visualization for Large High-dimensional Data Analysis.** to appear *in the Proceedings of the The 10th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid 2010), Melbourne, Australia, May 17-20 2010.*
- Seung-Hee Bae, Jong Youl Choi, et al. **Dimension Reduction Visualization of Large High-dimensional Data via Interpolation.** to appear *in the Proceedings of The ACM International Symposium on High Performance Distributed Computing (HPDC), Chicago, IL, June 20-25 2010*.

Questions?

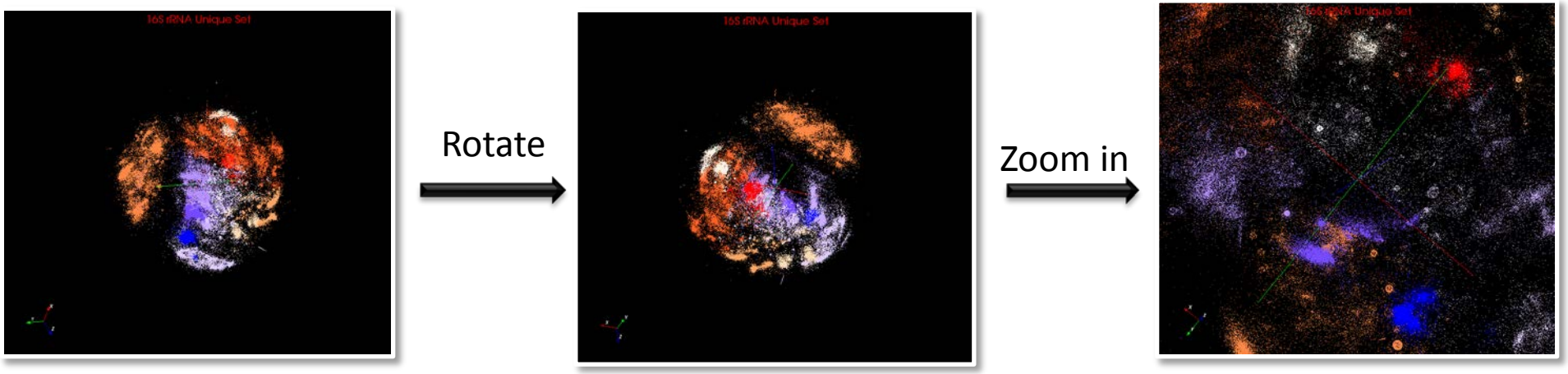
Backup Slides

DACIDR Flow Chart



Visualization

- Used PlotViz3 to visualize the 3D plot generated in previous step.
- It can show the sequence name, highlight interesting points, even remotely connect to HPC cluster and do dimension reduction and streaming back result.



All-Pair Sequence Analysis

- **Input: FASTA File**
- **Output: Dissimilarity Matrix**
- Use Smith Waterman alignment to perform local sequence alignment to determine similar regions between two nucleotide or protein sequences.
- Use percentage identity as similarity measurement.

```
ACATCCTTAACAA -  - ATTGC-ATC - AGT - CTA
| | | | | | | | | |         |  | | |     | |     |  |
ACATCCTTAGC - - GAATT - - TATGAT - CACCA
```

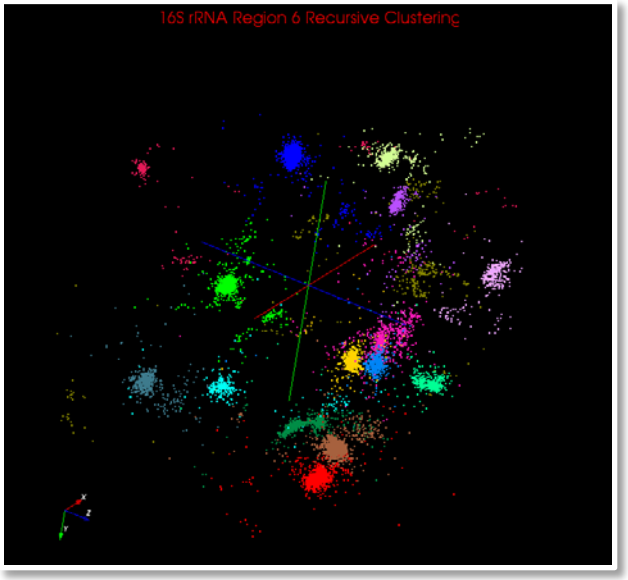
Deterministic Annealing

- Deterministic Annealing clustering is a robust pairwise clustering method.
- Temperature corresponds to pairwise distance scale and one starts at high temperature with all sequences in same cluster. As temperature is lowered one looks at finer distance scale and additional clusters are automatically detected.
- Multidimensional Scaling is a set of dimension reduction techniques. Scaling by Majorizing a Complicated Function (SMACOF) is a classic EM method and can be parallelized efficiently
- Adding temperature from DA can help prevent local optima problem.

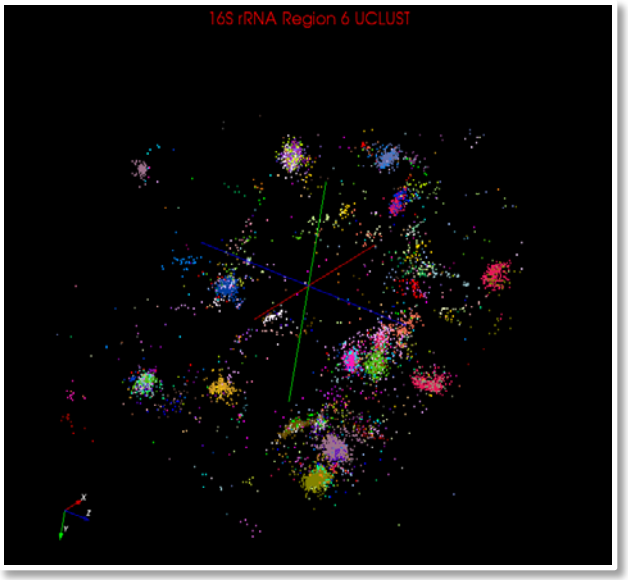
PWC vs UCLUST/CDHIT

	PWC	UCLUST					CDHIT		
Hard-cutoff Threshold	--	0.75	0.85	0.9	0.95	0.97	0.9	0.95	0.97
Number of A-clusters (number of clusters contains only one sequence)	16	6	23	71(10)	288(77)	618(208)	134(16)	375(95)	619(206)
Number of clusters uniquely identified	16	2	9	8	9	4	3	2	1
Number of shared A-clusters	0	4	2	1	0	0	0	0	0
Number of A-clusters in one V-cluster	0	0	12	62(10)	279(77)	614(208)	131(16)	373(95)	618(206)

PWC



UCLUST



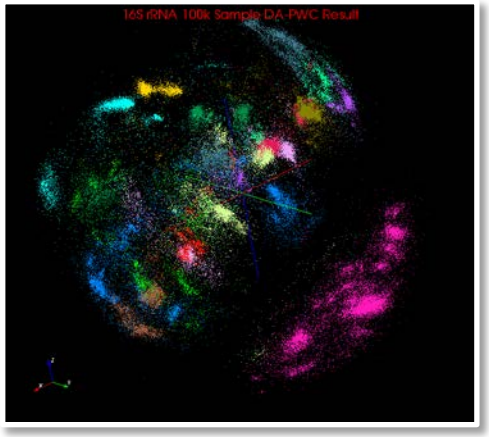
Heuristic Interpolation

- MI-MDS has to compare every out-sample point to every sample point to find k-NN points
- HI-MI compare with each center point of each tree node, and searches k-NN points from top to bottom
- HE-MI directly search nearest terminal node and find k-NN points within that node or its nearest nodes.
- Computation complexity
  - MI-MDS: $O(NM)$
  - HI-MI: $O(NlogM)$
  - HE-MI: $O(N(N_T + M_T))$

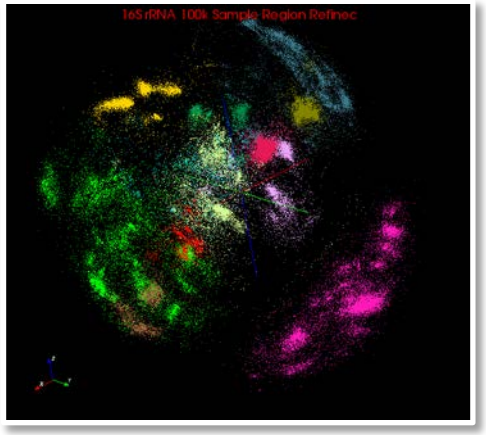
Region Refinement

- Terminal nodes can be divided into:
  - *V*: Inter-galactic void
  - *U*: Undecided node
  - *G*: Decided node
- Take a heuristic function *H(t)* to determine if a terminal node *t* should be assigned to *V*
- Take a fraction function *F(t)* to determine if a terminal node *t* should be assigned to *G.*
- Update center points of each terminal node *t* at the end of each iteration.

Before

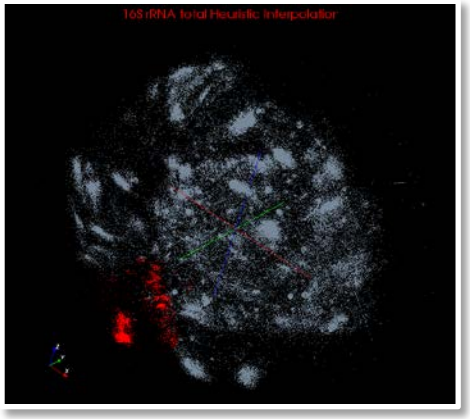


After

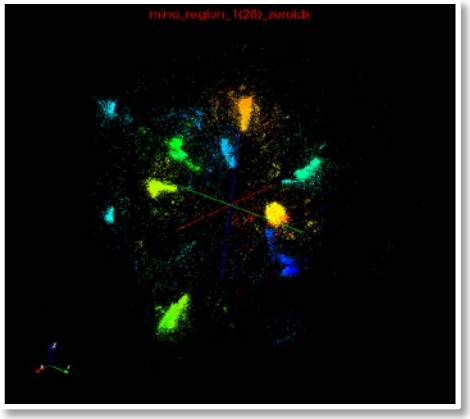


Recursive Clustering

- DACIDR create an initial clustering result $W = \{w_1, w_2, w_3, \ldots w_r\}$.
- Possible Interesting Structures inside each mega region.
- $w_1 \rightarrow W_1' = \{w1_1', w1_2', w1_3', \ldots, w1_{r_1}'\}$;
- $w_2 \rightarrow W_2' = \{w2_1', w2_2', w2_3', \ldots, w2_{r_2}'\}$;
- $w_3 \rightarrow W_3' = \{w3_1', w3_2', w3_3', \ldots, w3_{r_3}'\}$;
- …
- $w_r \rightarrow W_r' = \{wr_1', wr_2', wr_3', \ldots, wr_{r_r}'\}$;



Mega Region 1
Recursive
Clustering



Multidimensional Scaling

- **Input: Dissimilarity Matrix**
- **Output: Visualization Result (in 3D)**
- MDS is a set of techniques used in dimension reduction.
- Scaling by Majorizing a Complicated Function (SMACOF) is a fast EM method for distributed computing.
- DA introduce temperature into SMACOF which can eliminates the local optima problem.