Fast Parallel Construction of Correlation Similarity Matrices for Gene Co-Expression Networks on Multicore Clusters

Jorge González-Domínguez, María J. Martín

Computer Architecture Group, University of A Coruña, Spain
{jgonzalezd,mariam}@udc.es

International Conference on Computational Science
ICCS 2017

COMPUTER ARCHITECTURE GROUP
UNIVERSITY OF A CORUÑA
GAC.UDC.ES

1 Introduction

2 Parallel Construction of Similarity Matrices

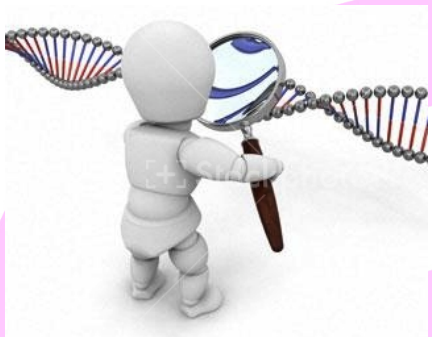
3 Experimental Results

4 Conclusions

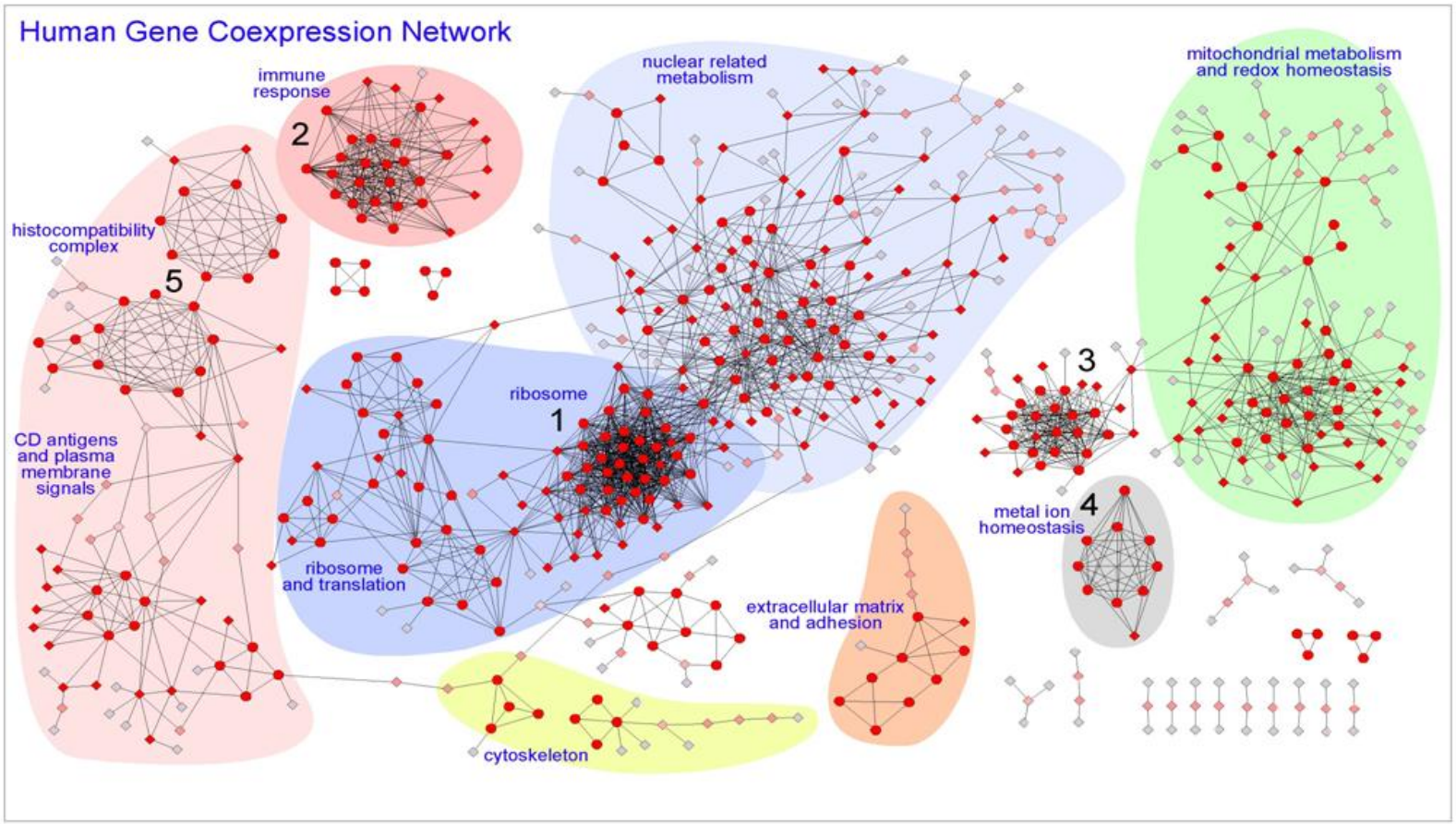
1 Introduction
2 Parallel Construction of Similarity Matrices
3 Experimental Results
4 Conclusions

Gene Co-Expression Networks (I)

- Graphical models to illustrate interactions among genes
- Connected groups of genes indicate biological relationships
  - Genes controlled by the same transcriptional regulatory program
  - Functionally related genes
  - Members of the same protein complex
  - More...
- Nodes represent genes.
- Edges represent interesting correlations.



Gene Co-Expression Networks (and II)



Calculation of Co-Expression Networks (I)

1. Read expression matrix
2. Construct similarity matrix
3. Calculate the threshold for the network
4. Construct the network discarding those elements lower than threshold

Calculation of Co-Expression Networks (II)

1. **Read expression matrix**
2. Construct similarity matrix
3. Calculate the threshold for the network
4. Construct the network discarding those elements lower than threshold

	S_1	S_2	S_3
G_1	43.26	40.89	5.05
G_2	166.6	41.87	136.65
G_3	12.53	39.55	42.09
G_4	28.77	191.92	236.56
G_5	114.7	79.7	99.76
G_6	119.1	80.57	114.59
G_7	118.9	156.69	186.95
G_8	3.76	2.48	136.78
G_9	32.73	11.99	118.8
G_{10}	17.46	56.11	21.41

Gene expression values

$\xrightarrow{|r(G_i, G_j)|}$ Pearson correlation

- The intensity of fluorescence in Microarrays or RNASeqs for each gene and sample
- Quantifies the expression of that gene in that sample

Calculation of Co-Expression Networks (III)

1. Read expression matrix
2. **Construct similarity matrix**
3. Calculate the threshold for the network
4. Construct the network discarding those elements lower than threshold

	G_1	G_2	G_3	G_4	G_5	G_6	G_7	G_8	G_9	G_{10}
G_1	1.00	0.23	0.61	0.71	0.03	0.35	**0.86**	**1.00**	**0.97**	0.37
G_2	0.23	1.00	0.63	0.52	**0.98**	**0.99**	0.29	0.30	0.46	**0.99**
G_3	0.61	0.63	1.00	**0.99**	0.77	0.53	**0.93**	0.56	0.41	0.51
G_4	0.71	0.52	**0.99**	1.00	0.69	0.41	**0.97**	0.66	0.52	0.40
G_5	0.03	**0.98**	0.77	0.69	1.00	**0.95**	0.48	0.09	0.27	**0.94**
G_6	0.35	**0.99**	0.53	0.41	**0.95**	1.00	0.17	0.41	0.57	**1.00**
G_7	0.86	0.29	**0.93**	**0.97**	0.48	0.17	1.00	**0.83**	0.72	0.16
G_8	**1.00**	0.30	0.56	0.66	0.09	0.41	0.83	1.00	**0.98**	0.42
G_9	**0.97**	0.46	0.41	0.52	0.27	0.57	0.72	**0.98**	1.00	0.58
G_{10}	0.37	**0.99**	0.51	0.40	**0.94**	**1.00**	0.16	0.42	0.58	1.00

Similarity (Co-expression) score

- Pearson's or other correlation measure for each gene pair

Calculation of Co-Expression Networks (IV)

1. Read expression matrix
2. Construct similarity matrix
3. **Calculate the threshold for the network**
4. Construct the network discarding those elements lower than threshold

	G_1	G_2	G_3	G_4	G_5	G_6	G_7	G_8	G_9	G_{10}
G_1	0	0	0	0	0	0	1	1	1	0
G_2	0	0	0	0	1	1	0	0	0	1
G_3	0	0	0	1	0	0	1	0	0	0
G_4	0	0	1	0	0	0	1	0	0	0
G_5	0	1	0	0	0	1	0	0	0	1
G_6	0	1	0	0	1	0	0	0	0	1
G_7	1	0	1	1	0	0	0	1	0	0
G_8	1	0	0	0	0	0	1	0	1	0
G_9	1	0	0	0	0	0	0	1	0	0
G_{10}	0	1	0	0	1	1	0	0	0	0

$|r(G_i, G_j)| >= 0.8$

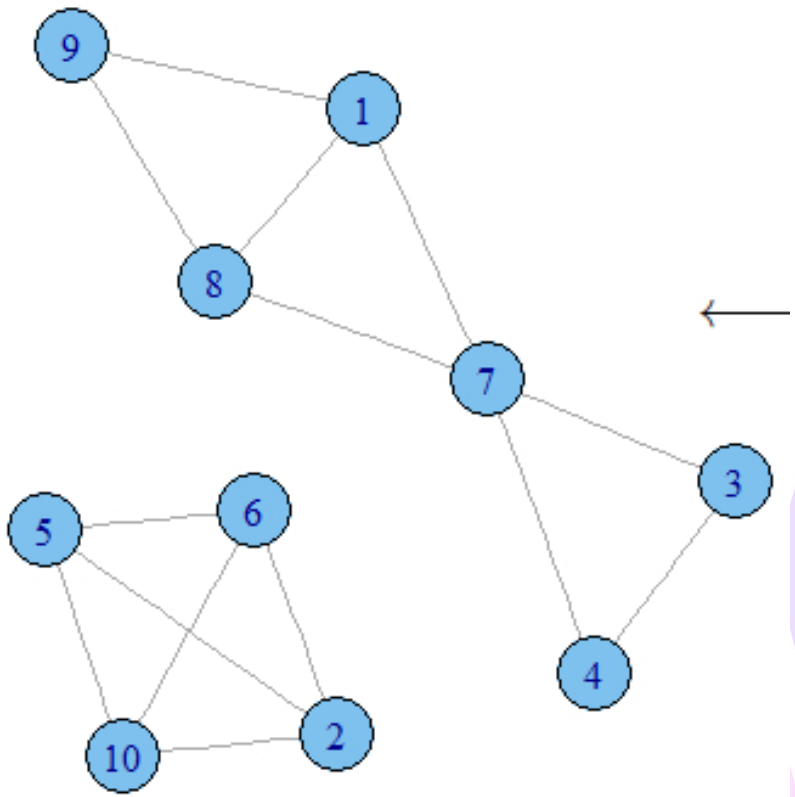
Significance threshold

Network adjacency matrix

- Based on the measures of the similarity matrices

Calculation of Co-Expression Networks (and V)

1. Read expression matrix
2. Construct similarity matrix
3. Calculate the threshold for the network
4. **Construct the network discarding those elements lower than threshold**



Background: *RMTGeneNet*

- *Scott M. Gibson, Stephen P. Ficklin, Sven Isaacson, Feng Luo, Frank A. Feltus, and Melissa C. Smith. Massive-Scale Gene Co-Expression Network Construction and Robustness Testing Using Random Matrix Theory. PLOS One, 8(2), 2013.*
- Three modules:
  - Pearson's correlation to construct similarity matrix
  - Random Matrix Theory (RMT) to calculate the threshold
  - Discard links with correlation value lower than threshold
- Networks with high robustness and sensitivity
- C++ implementation available at *https://github.com/spficklin/RMTGeneNet*

Goal of the work

- Module of *RMTGeneNet* to construct similarity matrices requires most of time
- **Acceleration of construction of similarity matrices with Pearson's correlation**
- **MPICorMat**
  - Improvement of memory accesses in the sequential computation
  - Exploitation of multicore clusters with MPI and OpenMP
  - Useful for large networks (Big Data)
  - It can substitute first module of *RMTGeneNet*
  - Available at *https://sourceforge.net/projects/mpicormat/*

1 Introduction

2 Parallel Construction of Similarity Matrices

3 Experimental Results

4 Conclusions

Programming technologies

MPI (Message Passing Interface)

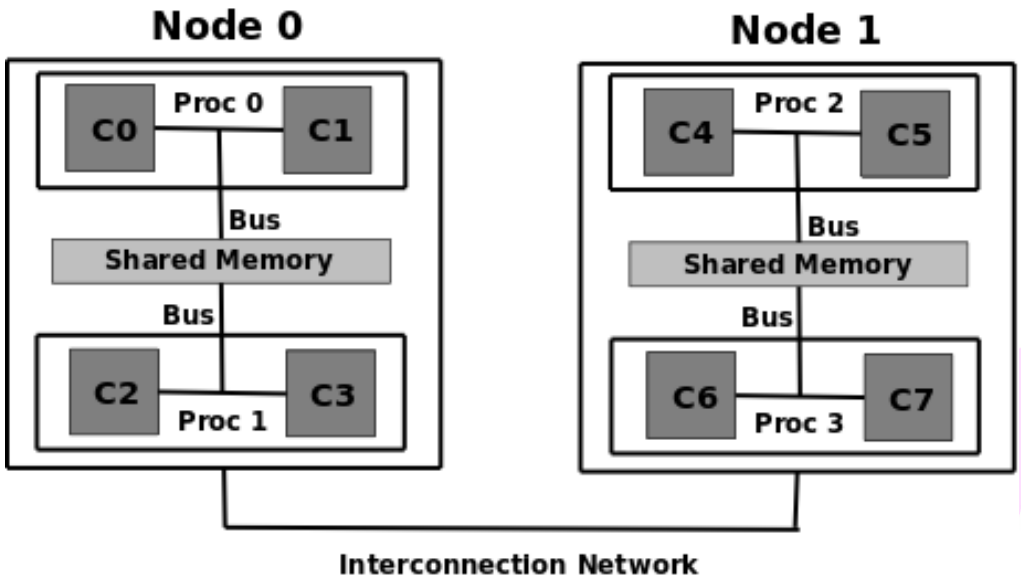
- *De-facto* standard for distributed memory systems
- Several processes with associated local memory
- Each process is associated to one core or a group of cores
- Data exchange performed through communication routines (often main performance bottleneck)

OpenMP

- Interface for shared memory systems
- A set of compiler directives inserted in the code
- Fork-join model: master thread creates slave threads that can perform different tasks

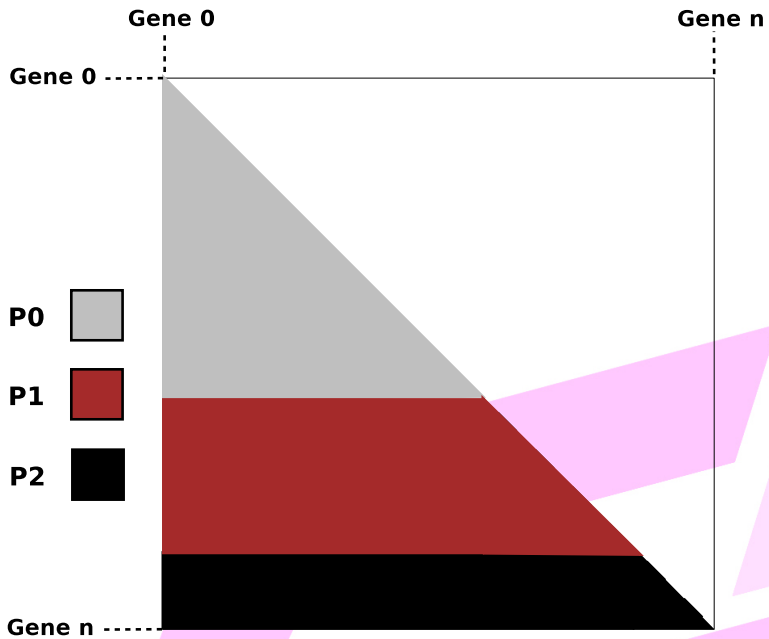
Data replication (and II)

2 nodes; 4 cores per node



Workload distribution

- All pairs (X,Y) with the same X to the same process
- Variable number of rows to balance the workload
  - Similar computational cost for each pair



Pseudocode of *MPICorMat*

1. Read input matrix M with the expression values;
2. Calculate *myIniRow* and *myLastRow*;
3. Initialize matrix of private scores $myS := 1$;
4. Initialize iterator $iter := 0$;
5. #pragma omp parallel for schedule(dynamic);
6. foreach row i from *myIniRow* to *myLastRow*{
   1. foreach column j from 0 to $i - 1$\{
      1. $myS[iter] := CalculatePearson(i, j)$; # GSL routine
      2. $iter++$;\}
   2. $iter++$;\} # Score for diagonal elements is 1.0;
7. Write partial result with *MPI_File_Write*(*myS*);

Data replication (I)

Advantage

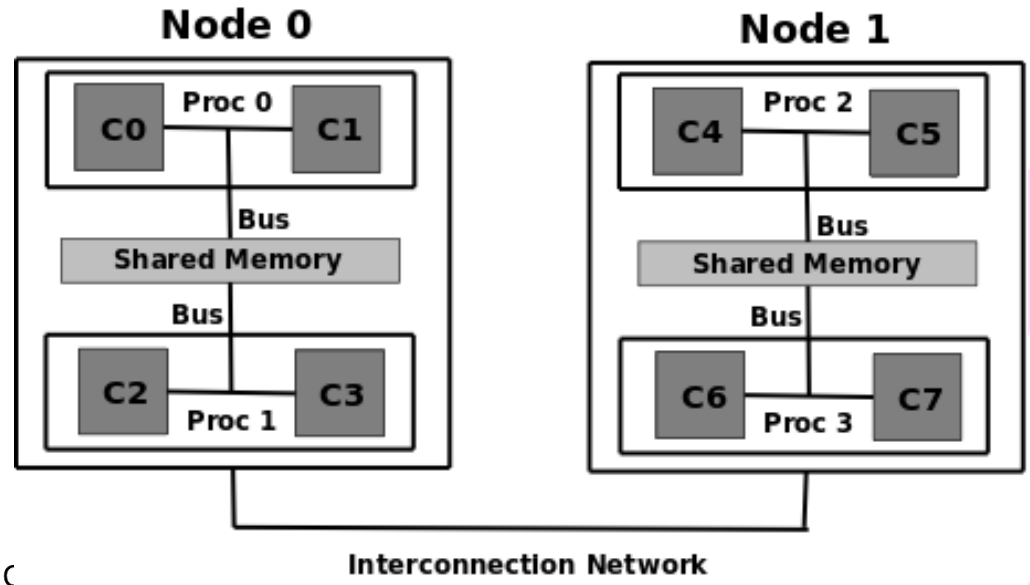
- All processes have their own copy of the expression matrix
- Communication avoidance: no communication during the matrix construction

Drawback

- Memory overhead
- We reduced it thanks to several threads working over the same copy of the matrix

Data replication (and II)

- Only MPI
  - One process per node
  - $M \times N \times 8$ floats
- MPI+OpenMP
  - One process per node, C threads per process
  - $M \times N \times 2$ floats



1 Introduction

2 Parallel Construction of Similarity Matrices

3 Experimental Results

4 Conclusions

System characteristics

Hardware

- 16 nodes connected through InfiniBand FDR
- Two 8-core Intel Xeon E5-2660 Sandy-Bridge processors per node (16 cores)
- Non Uniform Memory Access (NUMA) with 32MB per processor

Software

- OpenMPI v.1.7.2
- Support for OpenMP v.3.0
- GSL v.1.13 for Pearson's correlation

Datasets

Real data downloaded from the Geo Expression Omnibus (GEO) Dataset Browser available at the National Center for Biotechnology Information (NCBI) website

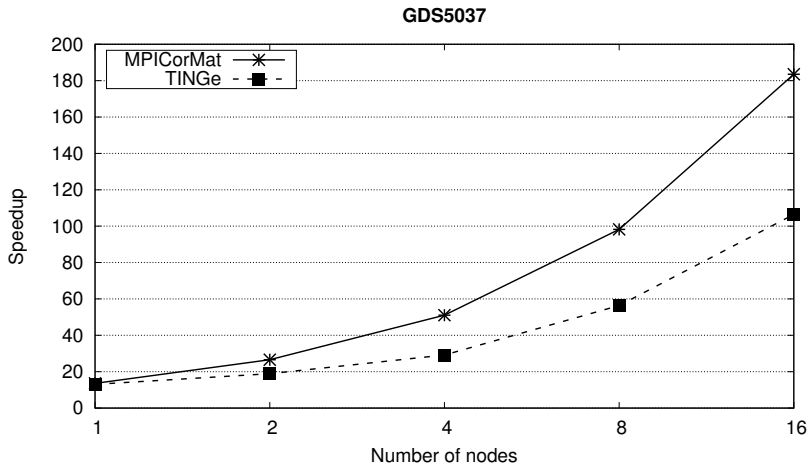
Name	Number of Genes	Number of Samples
GDS5037	41,000	108
GDS3242	61,170	128
GDS3244	61,170	160
GDS3795	54,675	200

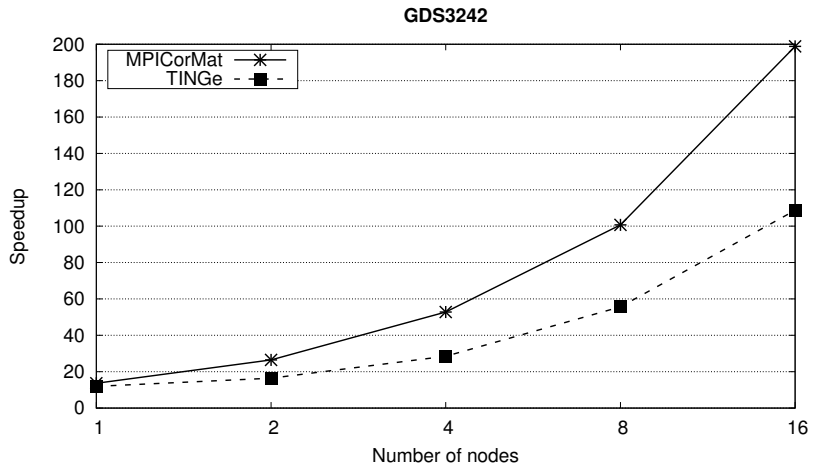
Summary of results (Runtime in seconds)

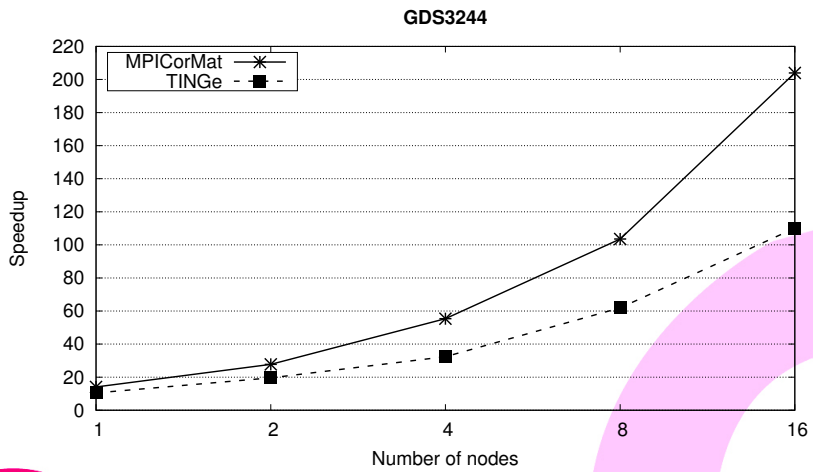
- *MPICorMat*: Two processes per node (one per processor) and eight threads per process. Pearson's correlation.
- *TINGe*: One process per core (no multithreading support). Mutual Information.

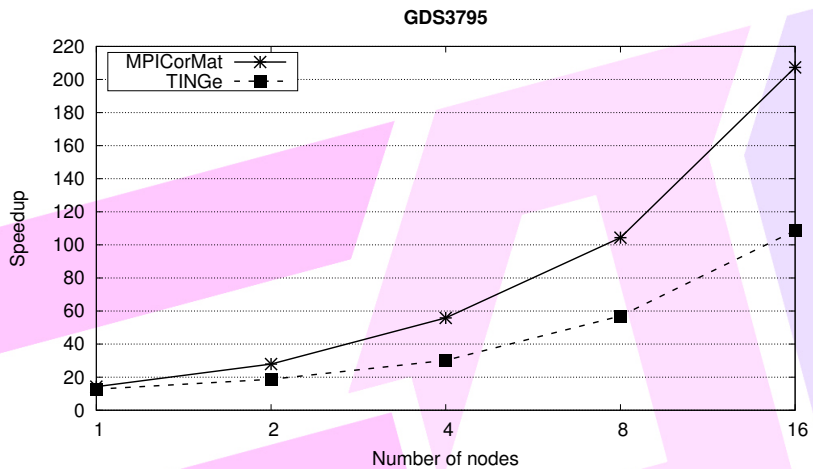
Cores	Tool	GDS5037	GDS3242	GDS3244	GDS3795
1	RMTGeneNet	5,336.51	13,124.76	16,004.83	15,470.96
	TINGe	5,206.48	12,442.99	14,664.41	12,965.41
	MPICorMat	2,539.96	6,652.75	8,365.14	8,139.81
16	TINGe	398.78	1,041.91	1,400.93	1,016.63
	MPICorMat	186.81	488.93	595.44	572.06
256	TINGe	48.91	114.47	129.71	119.35
	MPICorMat	13.84	33.46	41.02	39.26

Scalability results









1 Introduction

2 Parallel Construction of Similarity Matrices

3 Experimental Results

4 Conclusions

Summary

- *MPICorMat*, first tool to exploit multicore clusters to construct Pearson's correlation matrices
- Efficient hybrid MPI/OpenMP parallelization
- It can be used for the most expensive step in the generation of co-expression matrices
  - Instead of the first module of *RMTGeneNet*
  - Also useful in other fields
- Impressive speedups over the *RMTGeneNet* module
  - Around two times faster with the same resources (one core)
  - On average 390.43 times faster using 16 nodes.
- Faster and higher scalability than TINGe
- It will directly benefit from future GSL optimizations
- Available at *https://sourceforge.net/projects/mpicormat/*

Future Work

- Parallelization of the second *RMTGeneNet* module
  - Search of the RMT threshold
- Include support for additional correlation measures

Fast Parallel Construction of Correlation Similarity Matrices for Gene Co-Expression Networks on Multicore Clusters

Jorge González-Domínguez, María J. Martín

Computer Architecture Group, University of A Coruña, Spain
{jgonzalezd,mariam}@udc.es

International Conference on Computational Science
ICCS 2017

COMPUTER ARCHITECTURE GROUP
UNIVERSITY OF A CORUÑA
GAC.UDC.ES

UNIVERSIDADE DA CORUÑA