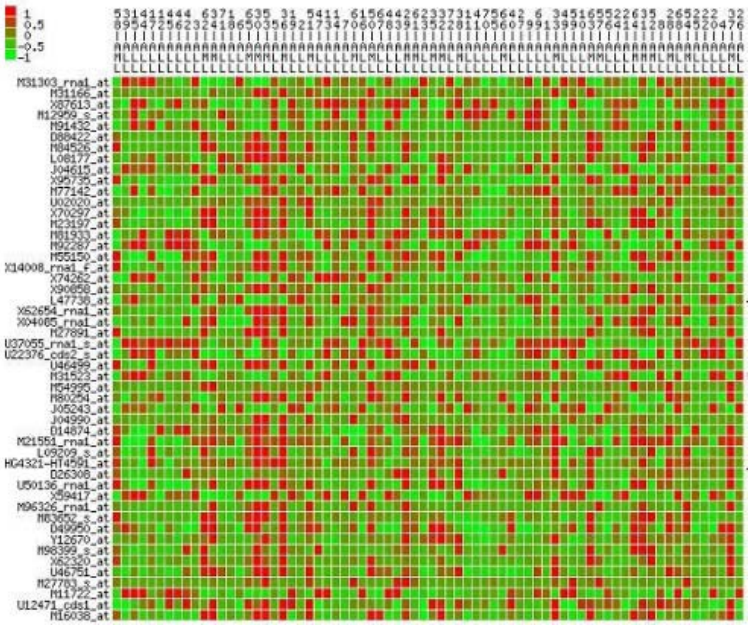
Construction Gene coexpression Network

CSE 633

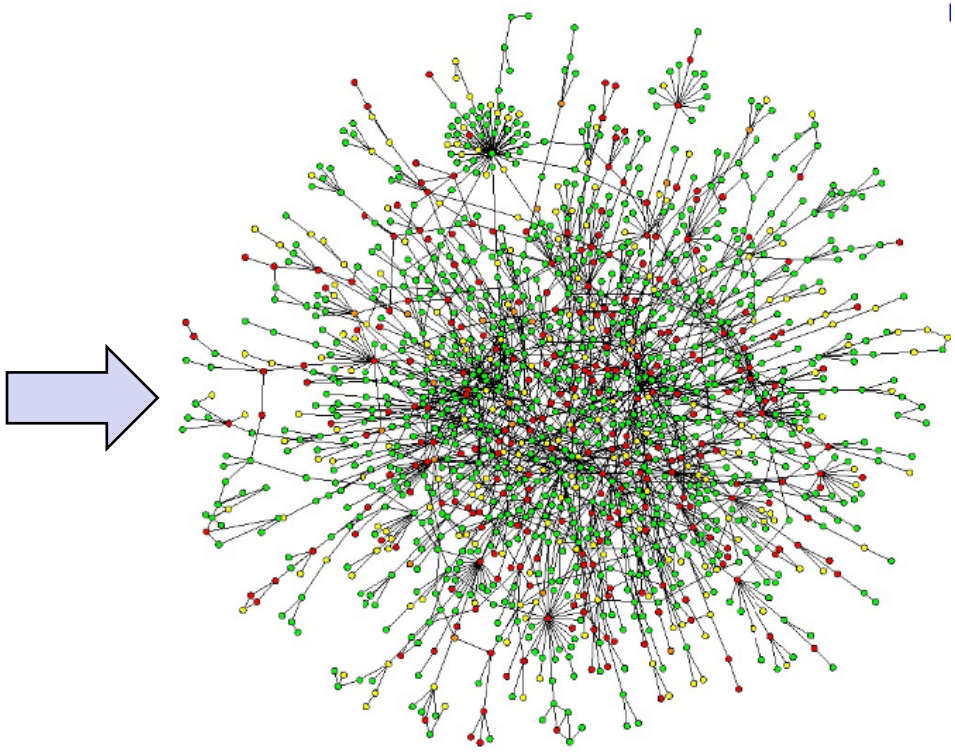
Nan Du

1 Dec 2011

The Problem



Gene Expression Data



Gene Coexpression Network

Process

1.Calculate the Correlation Coefficient between each gene pair

2. Eliminate the indirect interaction between genes

3. Keep Eliminating the edges between genes to meet the scale-free phenomenon

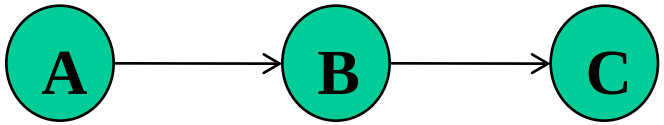
Step 1

- **Correlation Coefficient**

  - **A correlation coefficient indicates the extent to which two variables are related.**
  - **It can range from -1.0 to +1.0**
  - **A positive correlation coefficient indicates a positive relationship, a negative coefficient indicates an inverse relationship**
  - **Correlation CANNOT be equated with causality.**

Step 2

❖Remove the indirect influence between genes. Look at every triplet and remove the weakest link.



$$I(A,C) < \min[I(A,B), I(B,C)]$$

Step 3

A scale-free network is a network whose degree distribution follows a power law. That is, the fraction P(k) of nodes in the network having k connections to other nodes goes for large values of k as $P(k) \sim ck^{-\gamma}$

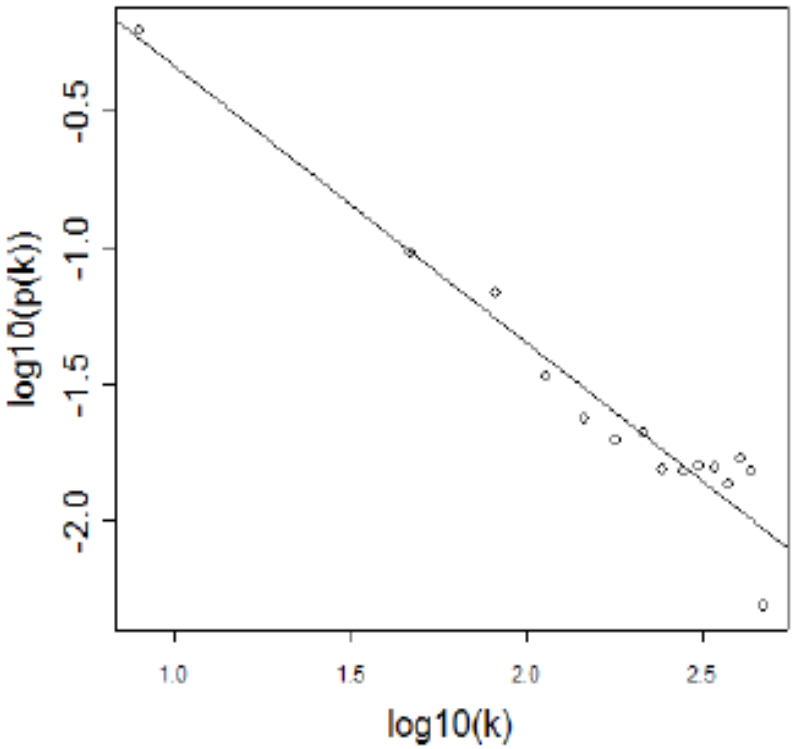
Gene Connectivity

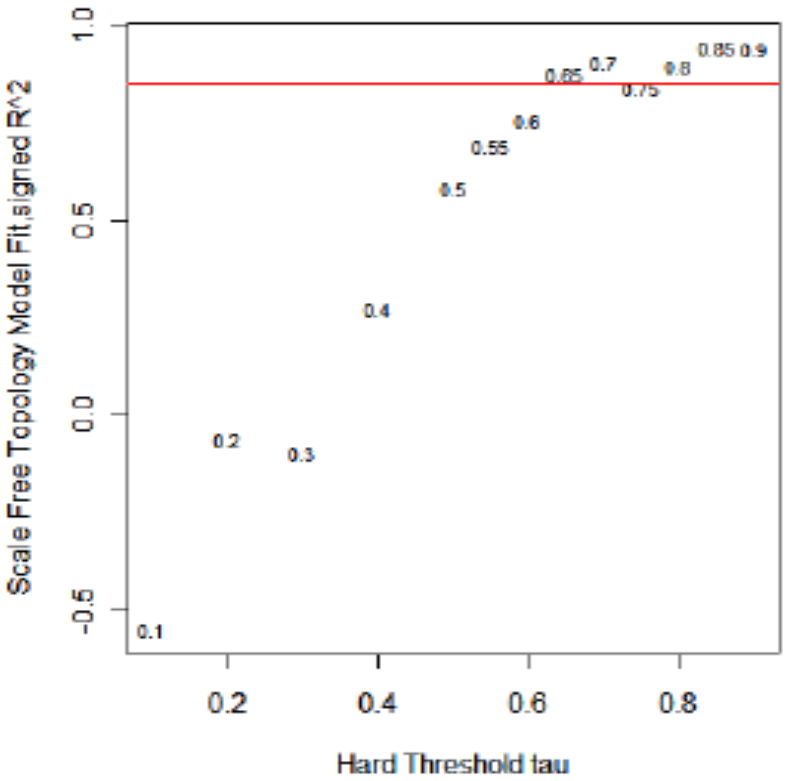
For unweighted networks=number of direct neighbors

For weighted networks= sum of connection strengths to other nodes

Then a fitting index R^2 is used to measure the scale-free topology degree which is the correlation between log(p(k)) and log(k)(where $P(k)$ notes the fraction of nodes in the network having k connections to other nodes goes for large values of k) . If R^2 of the model approaches 1, then there is a straight line relationship between log(p(k)) and log(k).

Process 3


log10(p(k))
-0.5
-1.0
-1.5
-2.0
1.0
1.5
2.0
2.5
log10(k)

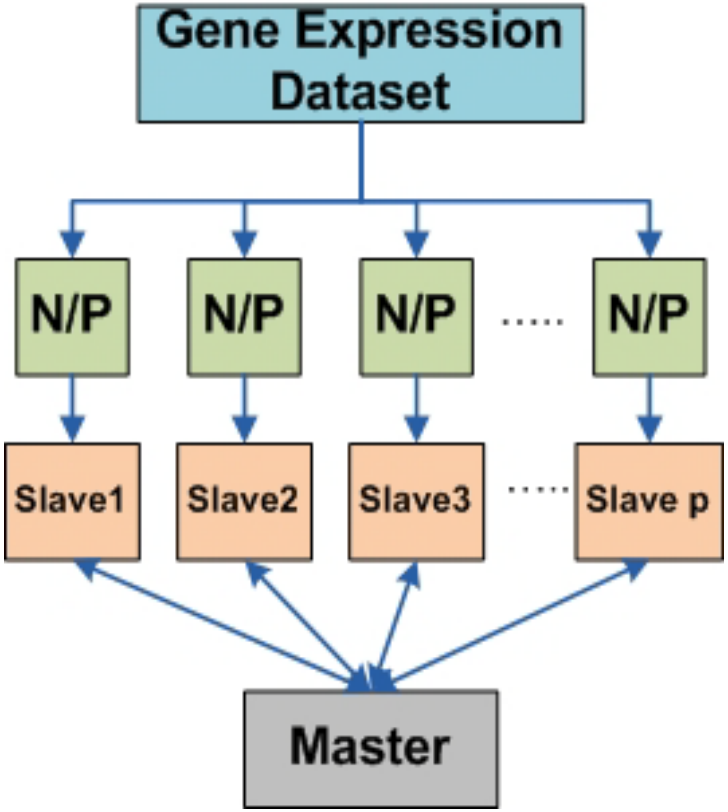

Scale Free Topology Model Fit,signed R^2
1.0
0.5
0.0
-0.5
0.1
0.2
0.3
0.4
0.5
0.55
0.6
0.65
0.7
0.75
0.8
0.85
0.9
0.2
0.4
0.6
0.8
Hard Threshold tau

Why Parallel?

Our algorithm's complexity is $O(N^2M^2)$, where N is the number of genes and M is the number of samples.

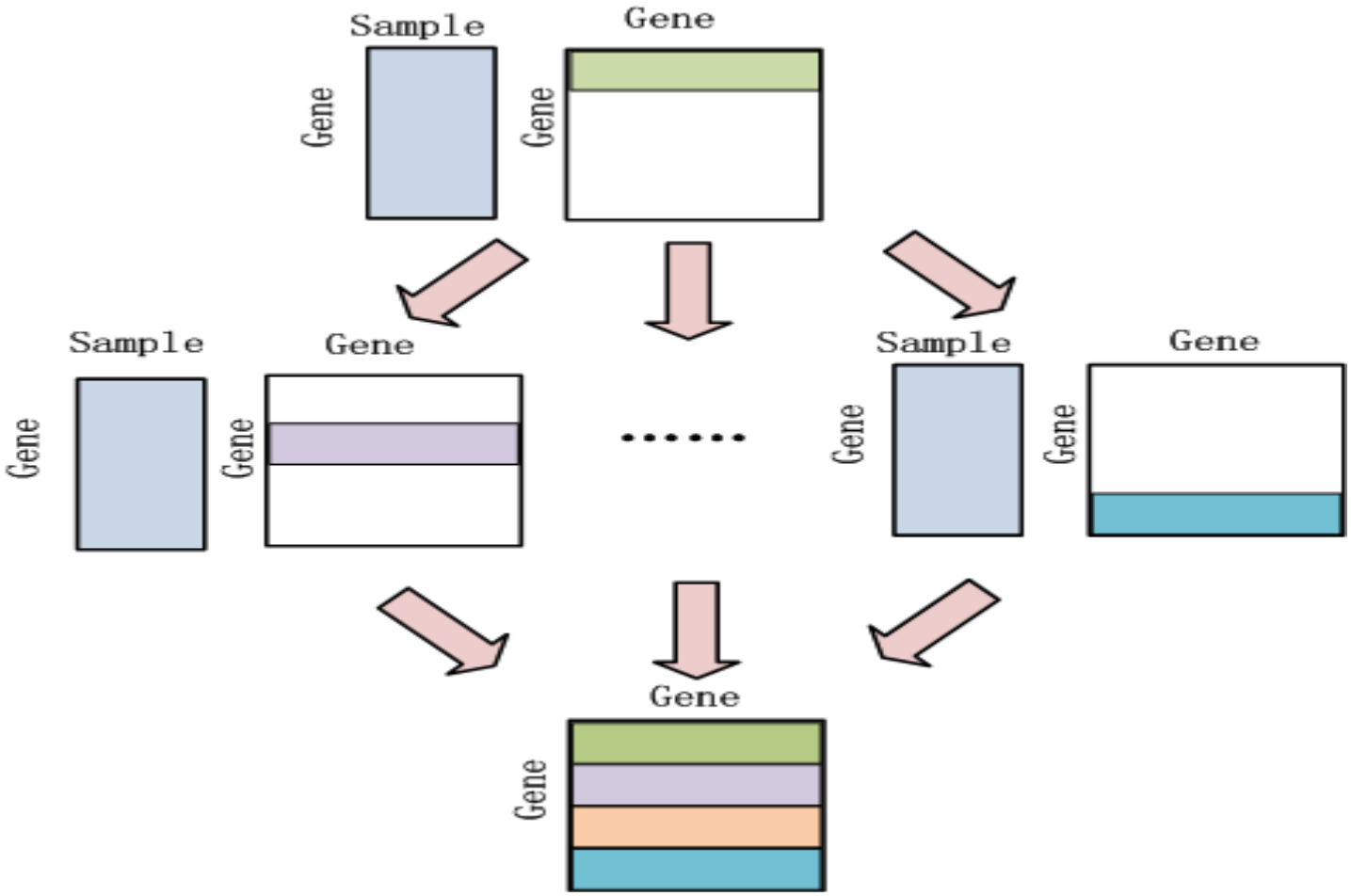
In our case, M is a constant which equals to 24

Parallel Solution



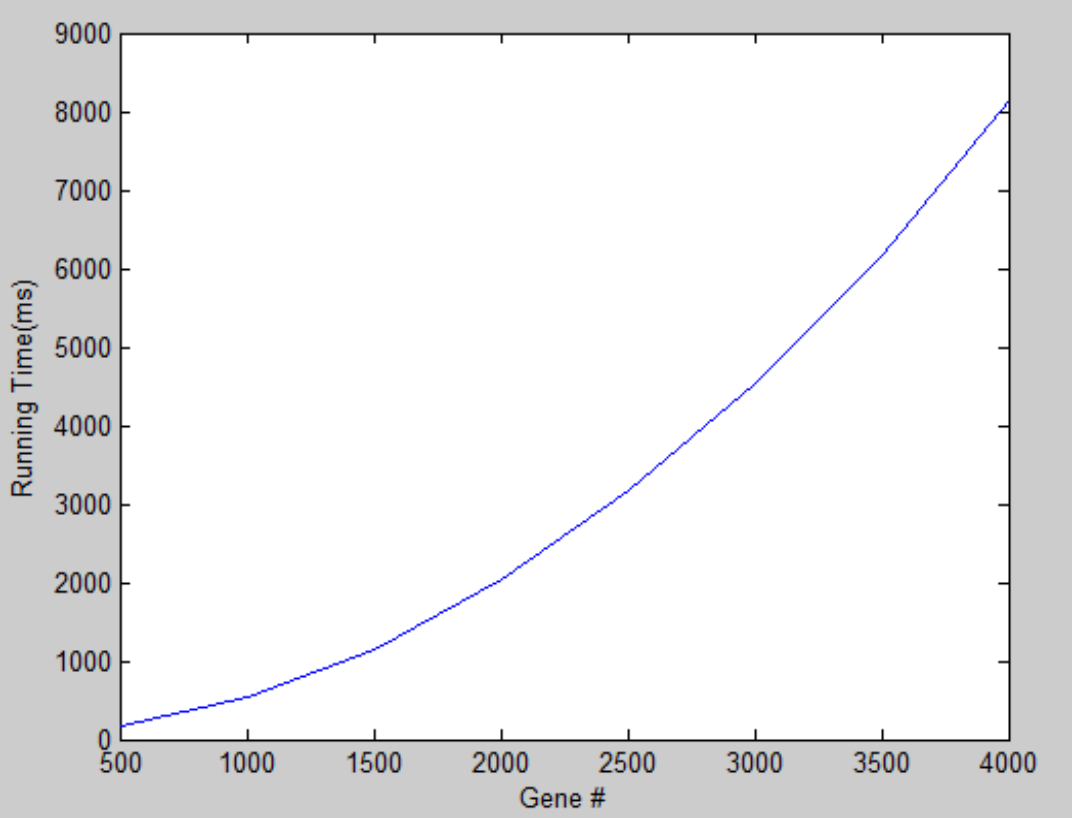
- Assign each processor with the whole data.
- In each slave processor, calculate parts of the Correlation Coefficient and output an array as the result.
- The master processor will gather the results and performs sequential computations.

Parallel Solution



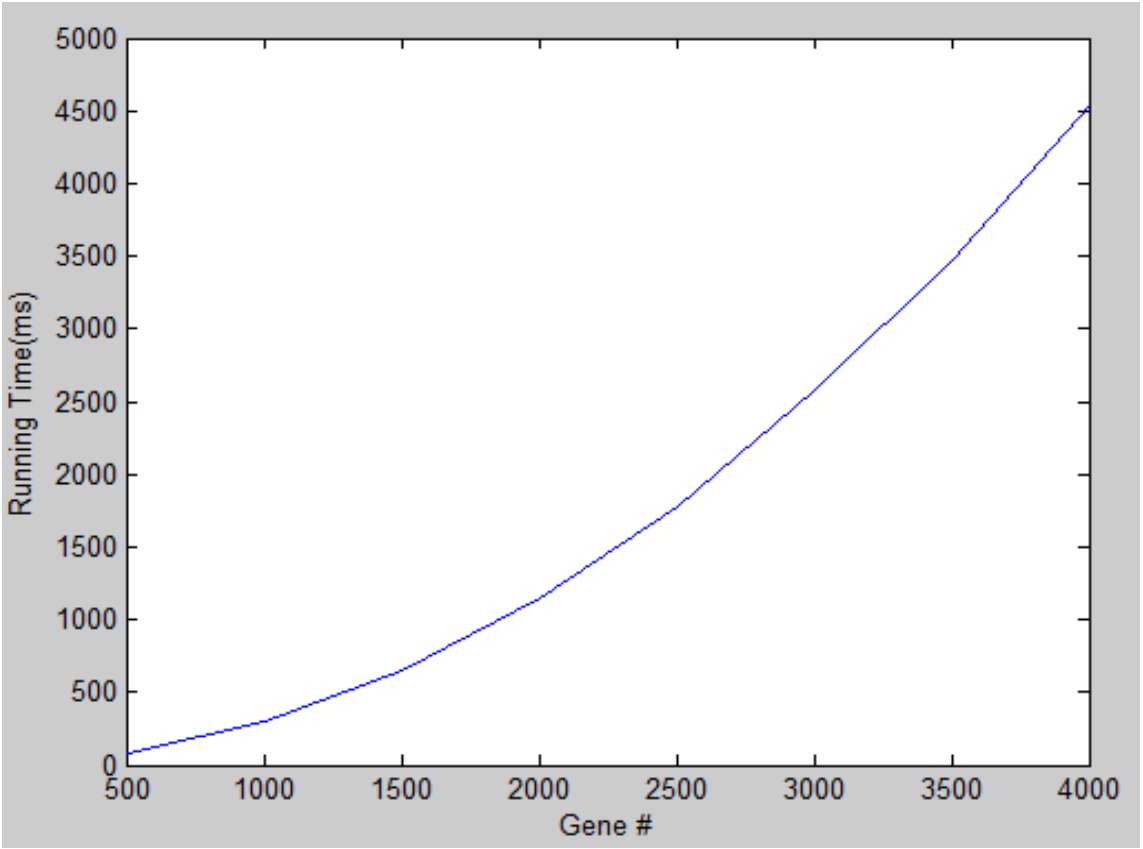
Result

Total running time
When nodes=1 ppn=2 under different data size



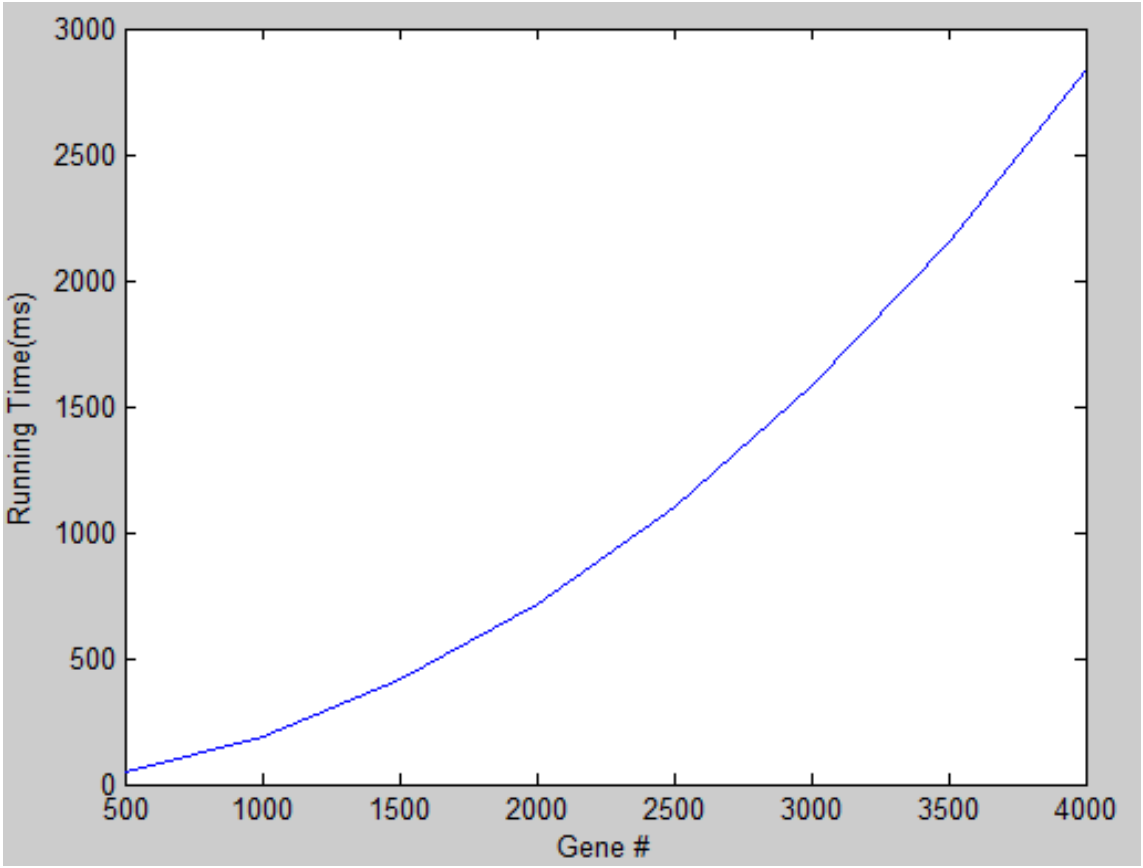
Result

Total running time and speedup
When nodes=2 ppn=2 under different data size



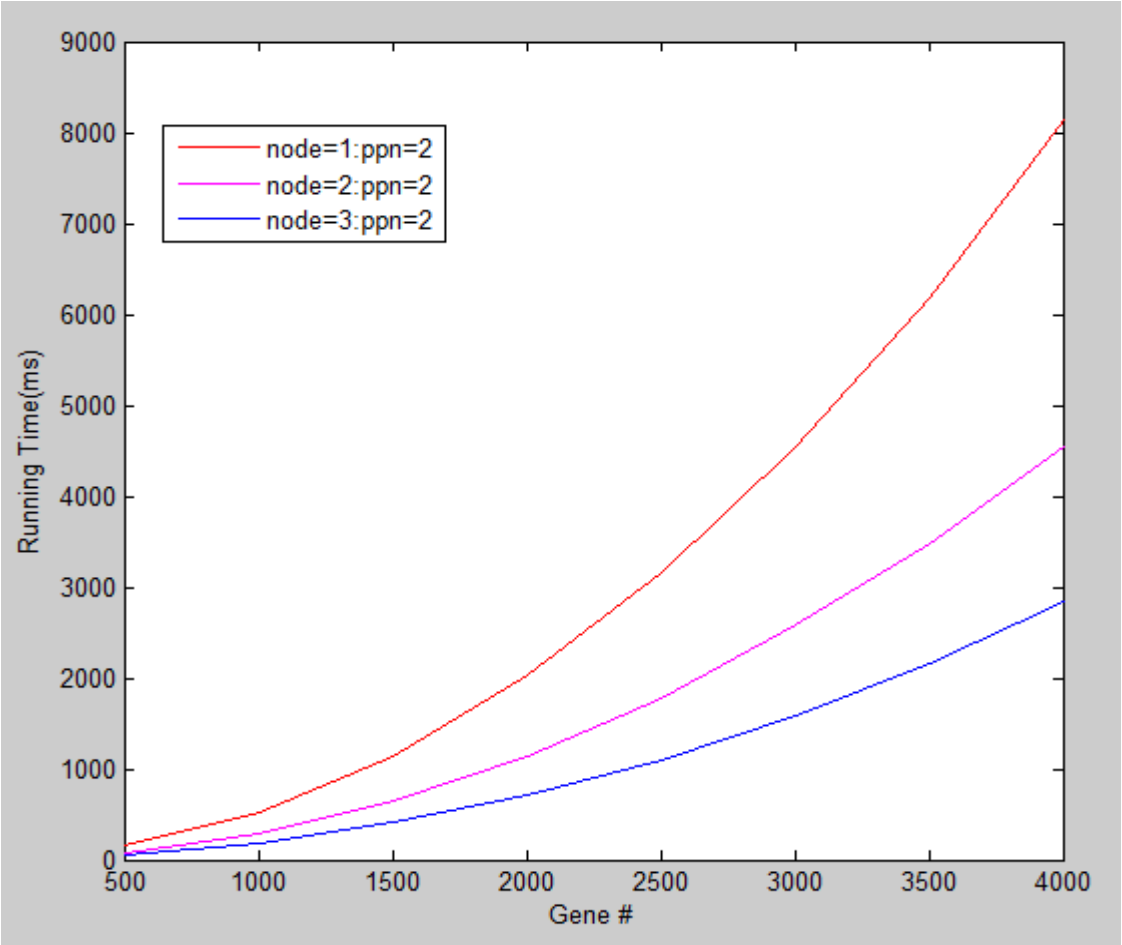
Result

Total running time and speedup
When nodes=3 ppn=2 under different data size



Result

Show them together



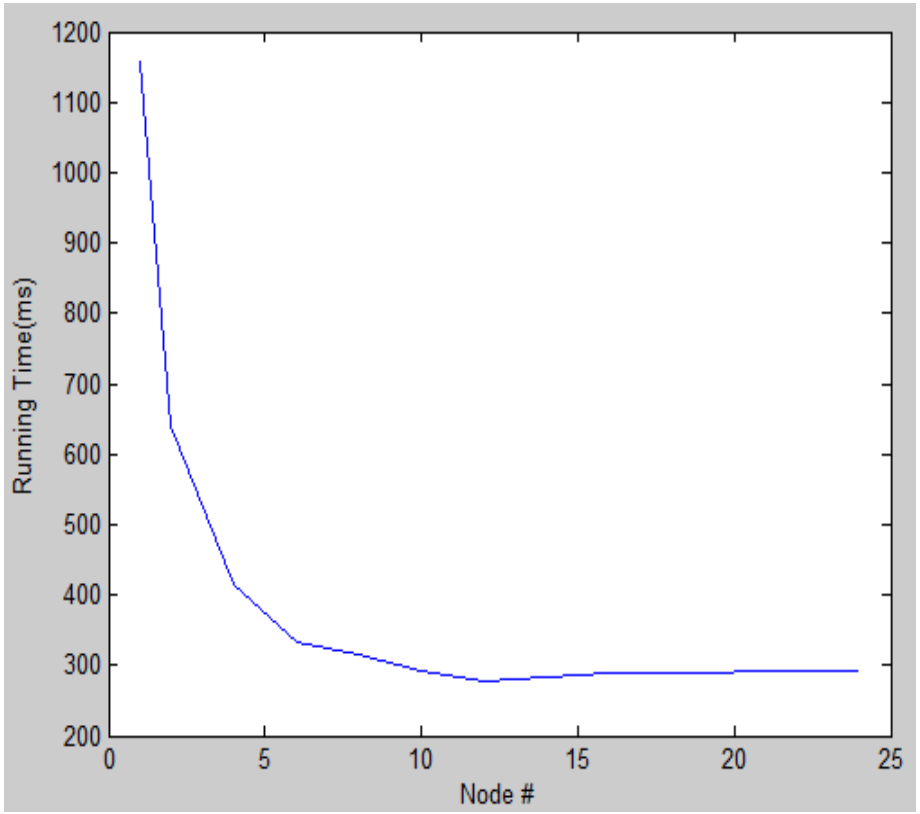
Result

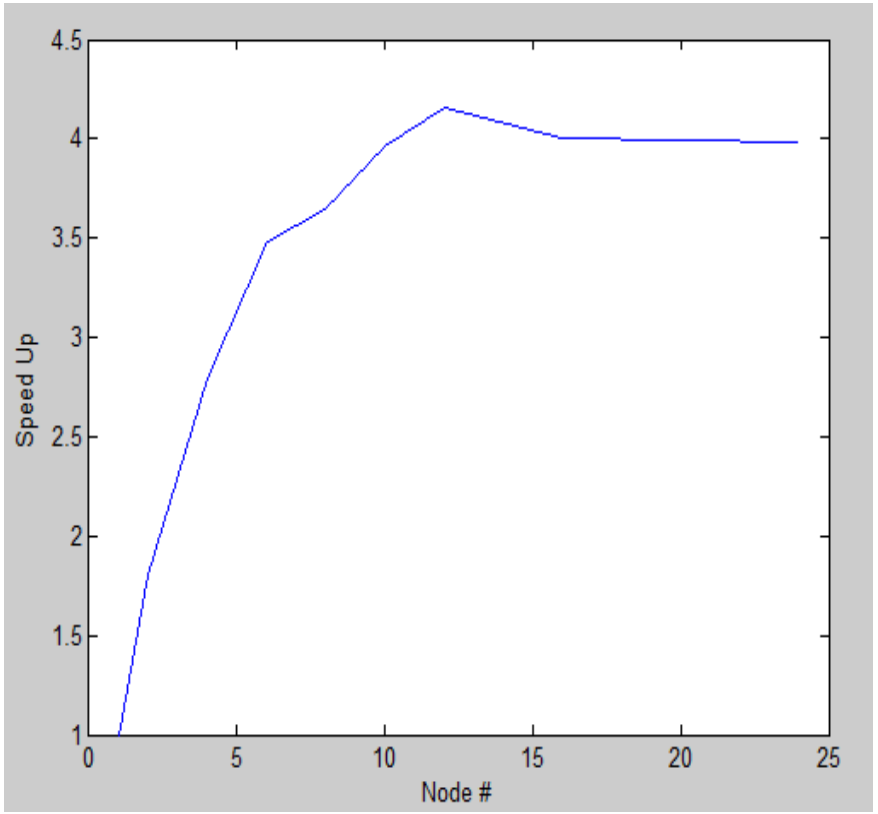
When we fix the # Gene as 1500

Node #	The # Gene for Each Node
1	1500
2	750
4	375
6	250
8	188
10	150
12	125
16	94
24	63

Result

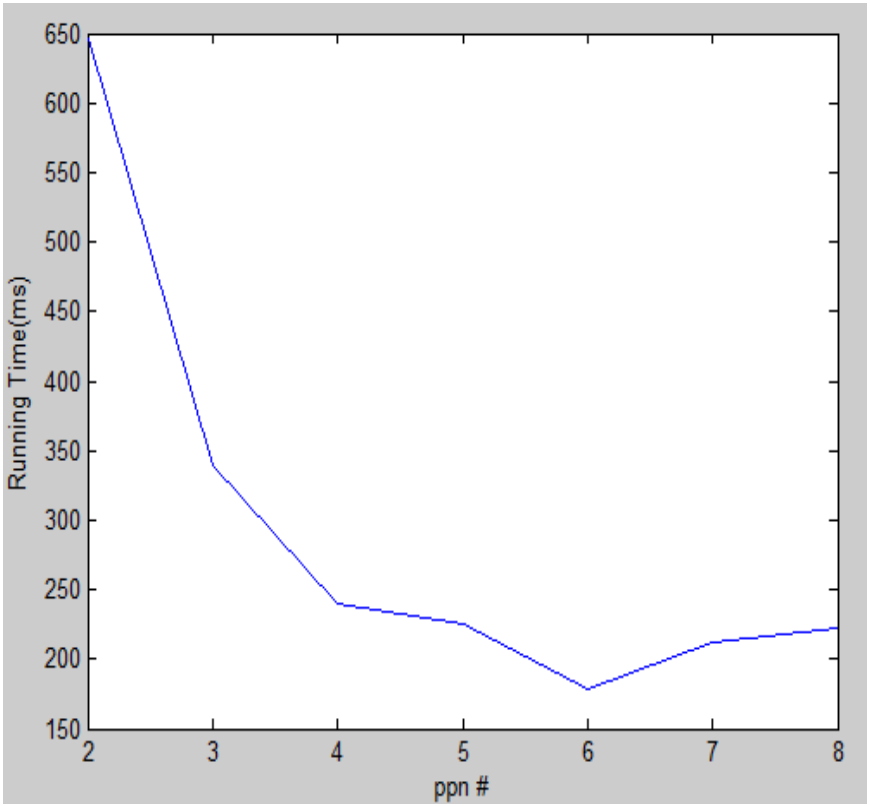
Total running time and speedup
Fixed the gene#=1500 and ppn=2,change the # node

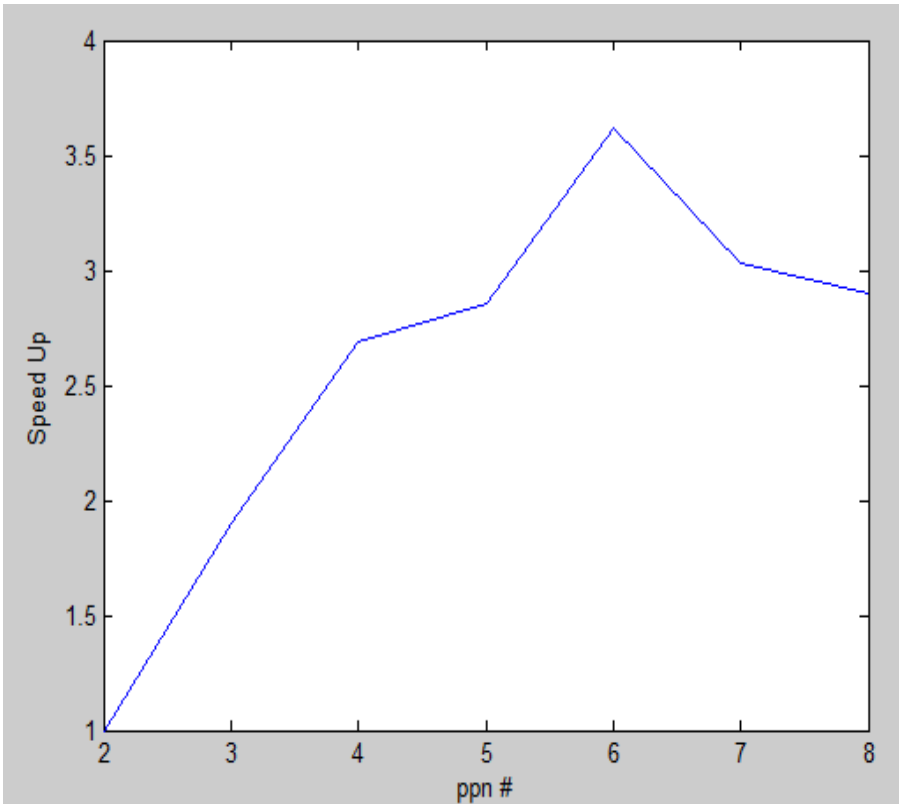




Result

Total running time and speedup
Fixed the gene#=1500 node#=2,change the ppn

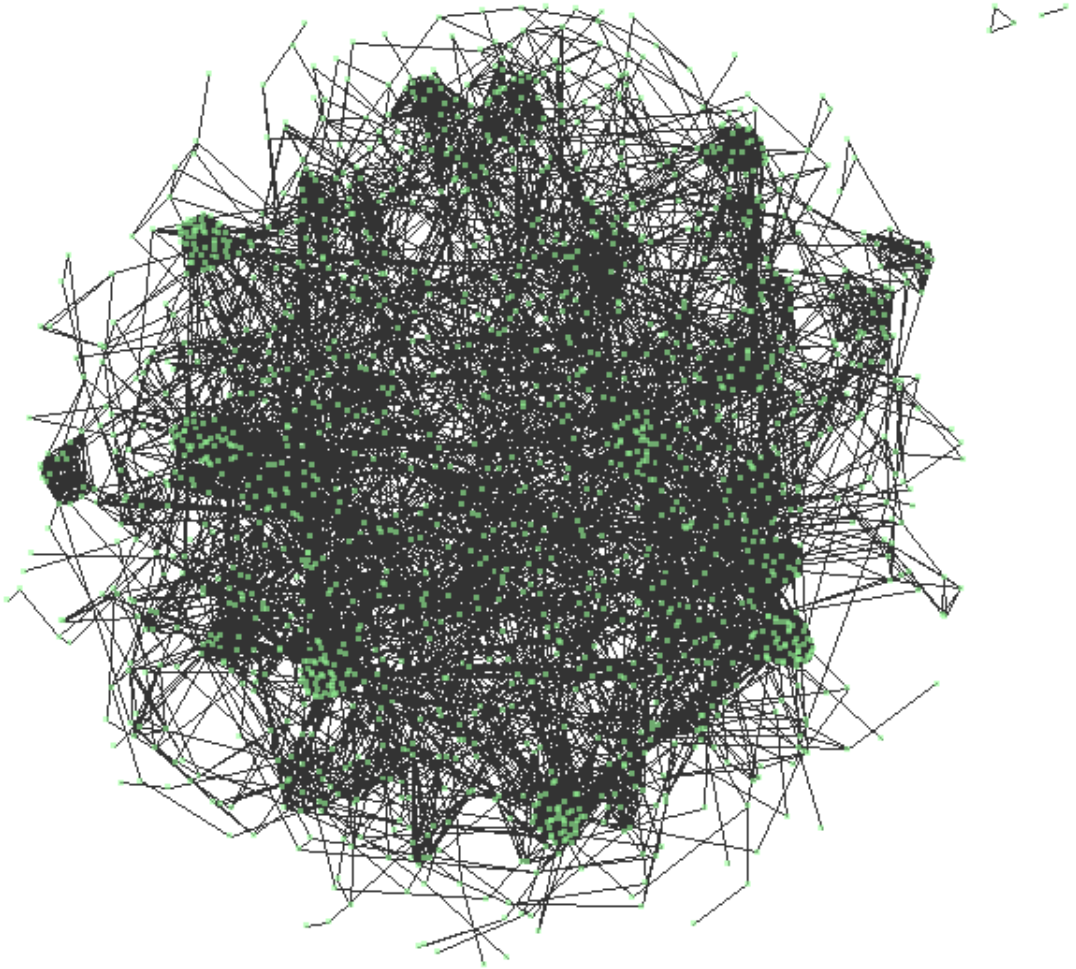




Result

# Node	ppn	# Core	Time
2	4	8	0.240237
4	2	8	0.413974
2	6	12	0.178385
6	2	12	0.332451
2	8	16	0.222706
8	2	16	0.316832

Result



Questions?

Thank you