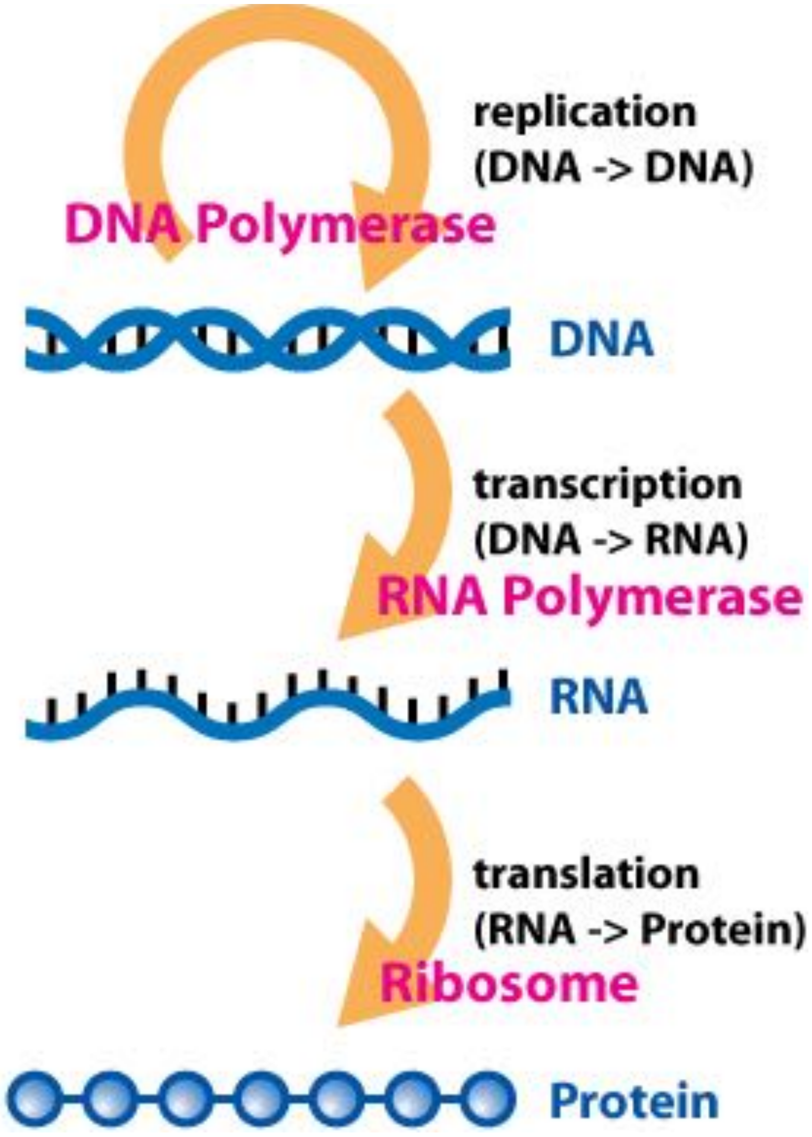
Discovery of Localization motifs in 3’UTRs using Biological Networks

By: Qasim Ahmed
Supervised by: Dr. Mathieu Lavallée-Adam
Department of Biochemistry, MicroBiology, and Immunology
Co-Supervised by : Dr. Marcel Turcotte
School of Information Technology and Engineering

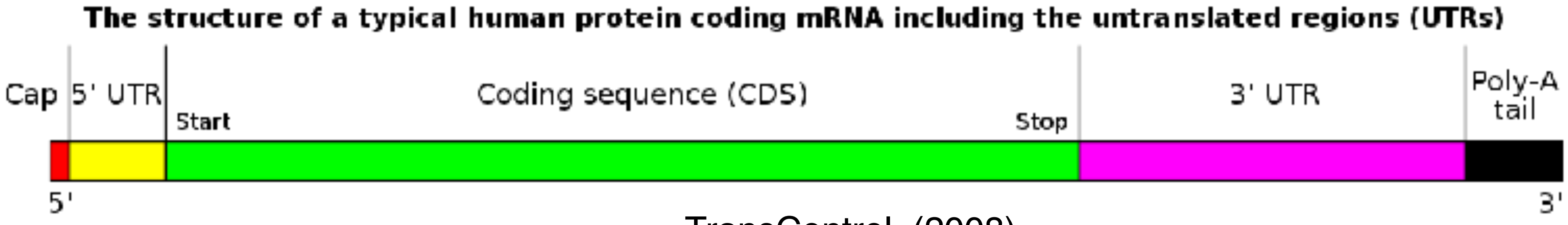
Central Dogma of Biochemistry



An overview of the (basic) central dogma of molecular biochemistry with all enzymes labeled.
Horspool, D.A. (2008).

3’ UTR

- Three prime untranslated region (3’UTR) is the section of messenger RNA that immediately follows the translation termination codon.
- Involved in the fine tuning of protein production


The structure of a typical human protein coding mRNA including the untranslated regions (UTRs)
Cap
5' UTR
Coding sequence (CDS)
Start
Stop
3' UTR
Poly-A tail
5'
3'

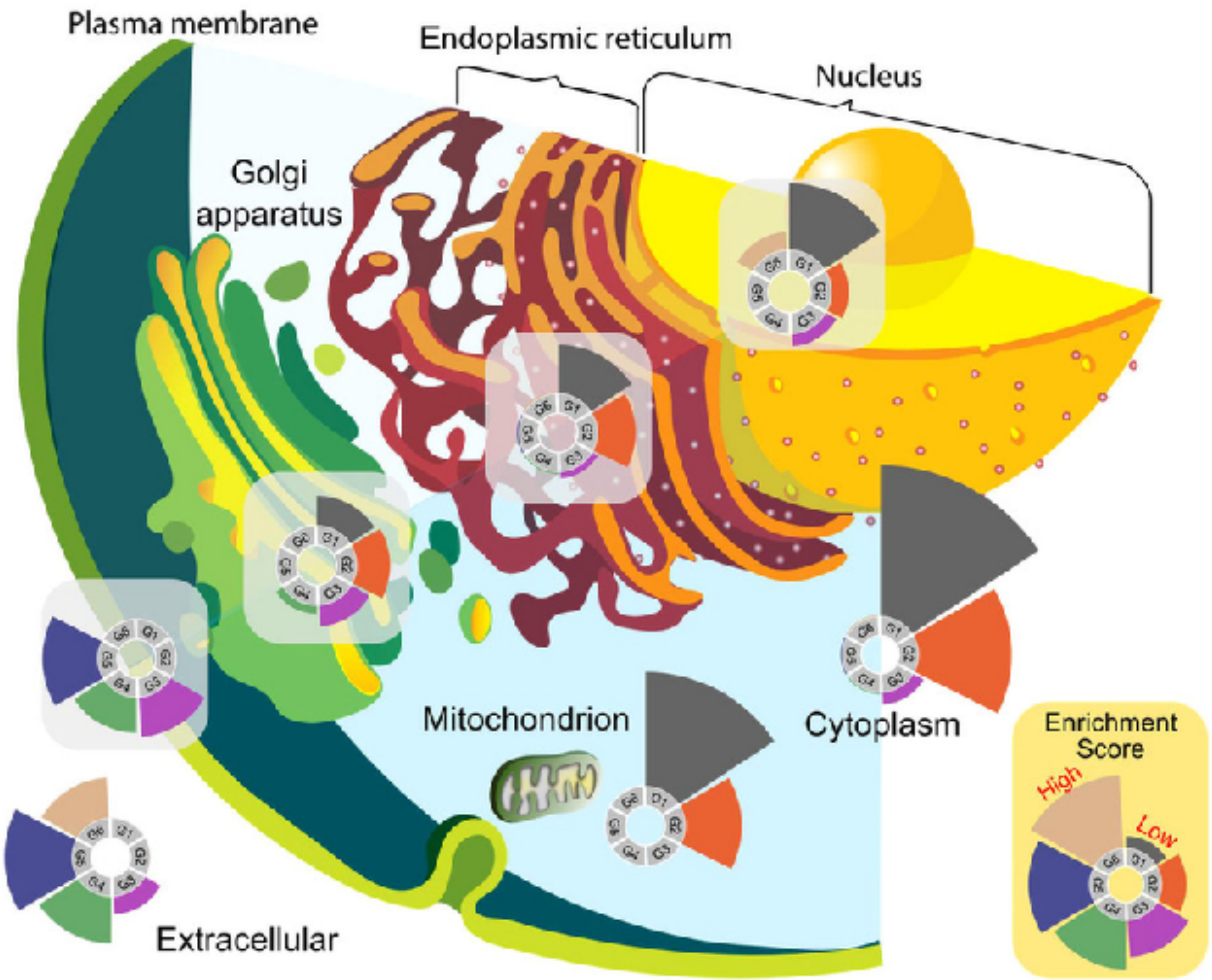
TransControl. (2008)

RNA Sequence motifs

- “Sequence motifs are short, recurring patterns in DNA that are presumed to have a biological function.”(D’haeseleer, 2006)

- Localization motifs

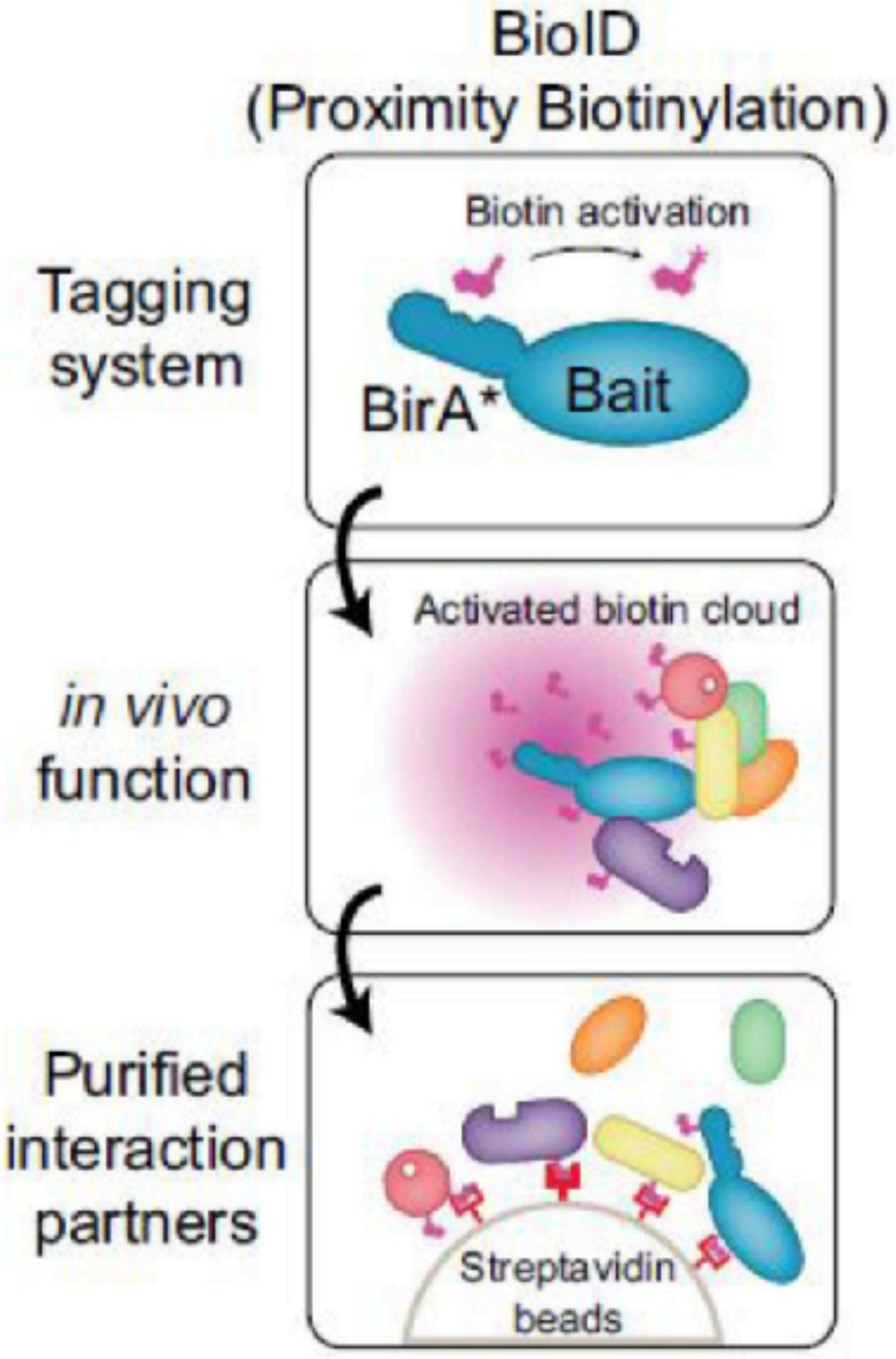
Localization motifs



The subcellular localization and age group composition of human proteins.
Chen, C.C. (2014).

BioID

- “BioID is a **proximity** biotinylation approach for mapping protein-protein interactions for chromatin-associated proteins.” (Jean-Philippe Lambert et. al, 2015)



Running Multiple BioID Experiments

	MDH2	OAT	PPIF	...
Experiment 1	762	876	2158	
Experiment 2	471	1658	346	
Experiment 3	1259	1181	582	
...				

Constructing Biological Network

	MDH2	OAT	PPIF
MDH2	1	-0.47	-0.03
OAT	-0.47	1	-0.86
PPIF	-0.03	-0.86	1

***Pearson Correlation Coefficients**

Sample BioID protein network

Problem

- The biological network stores correlation value between each proteins.
- How do we meaningfully derive a correlation value between two sets of proteins?

Graph based approach

- We can think of the biological network in terms of a graph G(E,V) where (V = proteins, E = correlation values)
- In this graph we explore two different scoring measures to approximate the correlation between all proteins in a set.

Method 1: Average

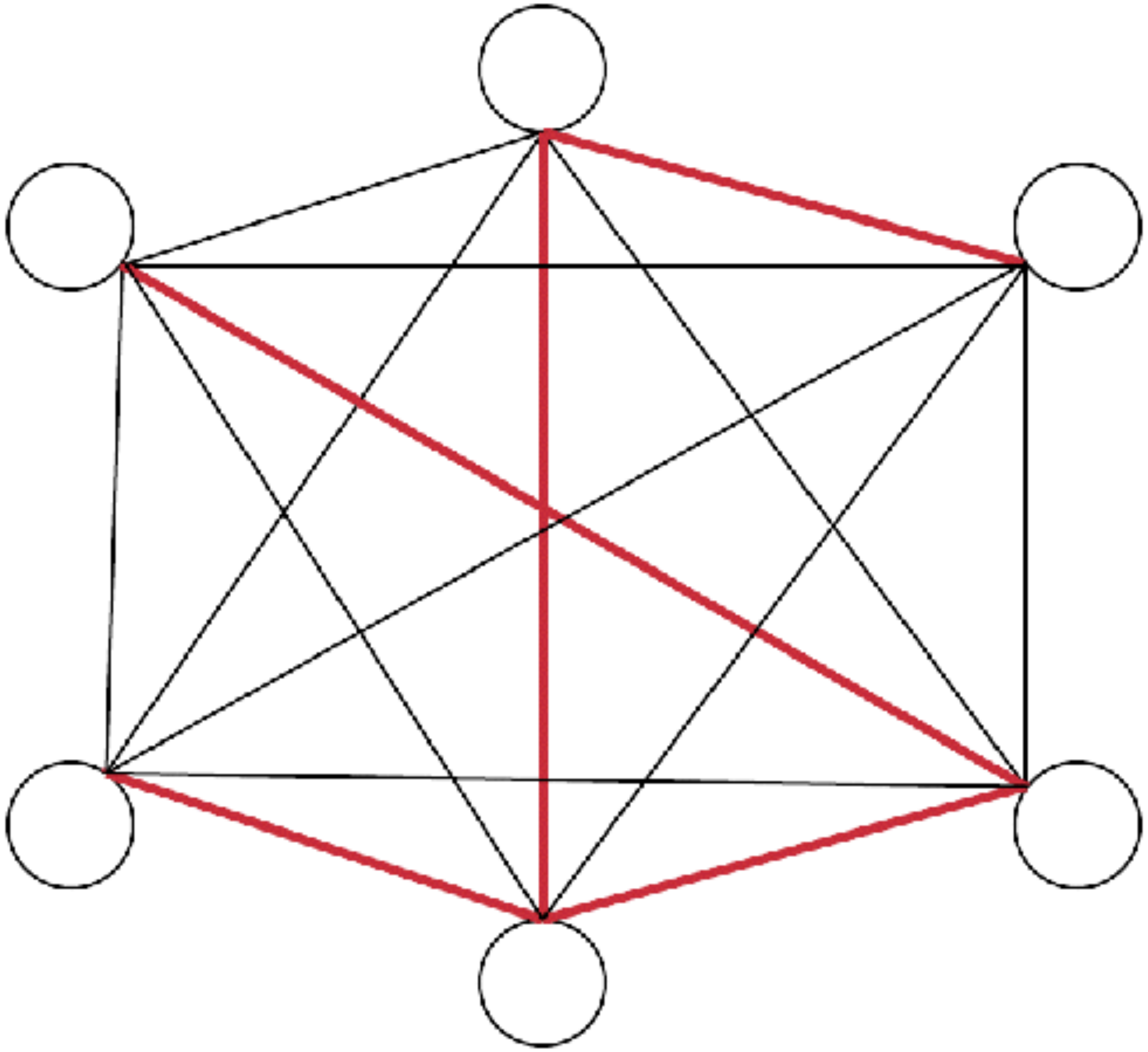
- Trivially take the average correlation value, between all edges.

- Bad (too much noise)

Method 2: Max Spanning Tree

- Modify Prim’s algorithm to find maximum spanning tree.

Reducing the noise



Methodology

- Generating all 7^8 motifs
  (all words of size 8 from an alphabet of {a c g u [ag] [cu] *})

- For each motif:

  1. Find the set of associated proteins.

  2. Score the correlation of that set of proteins.

  3. Determine the significance of that score

- Return all motifs with a significant score

Measuring Significance

- With our null hypothesis being non localization, we define significance as follows:

- Let P_k denote set of randomly selected proteins from the network (without replacements)

$$p - value(m_i) = prob(score(P_k) > score(P_i)) : |P_i| = |P_k|$$

Approximating p-values

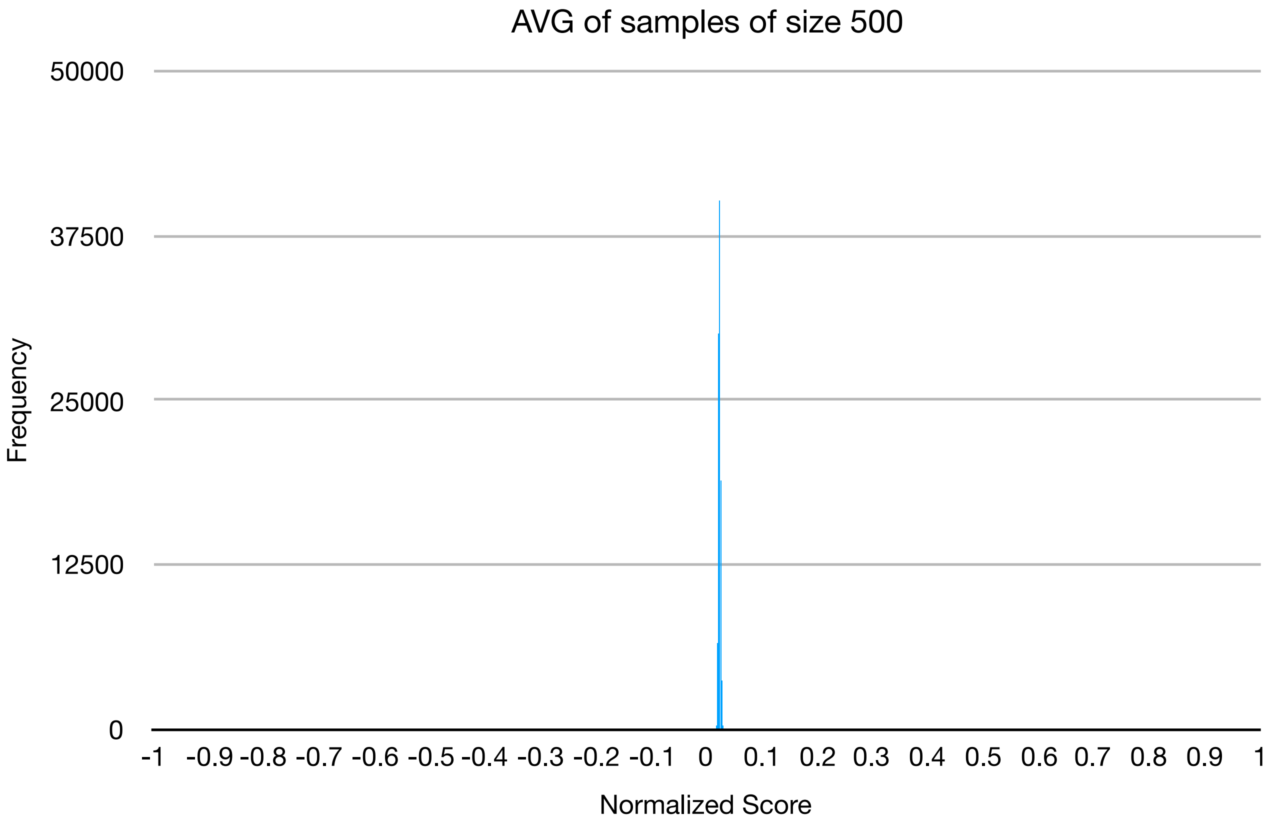
- To get the true probability would be unfeasible.

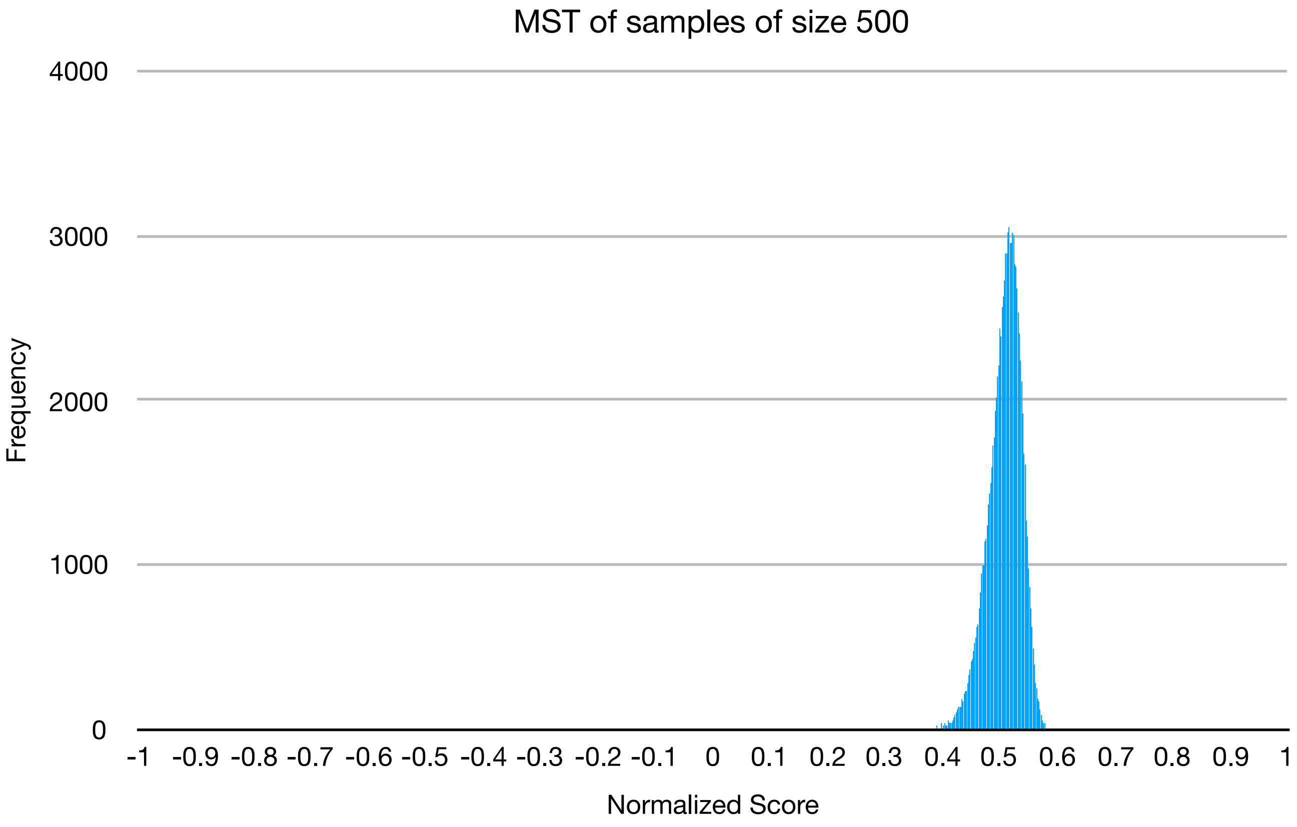
- We follow a Monte Carlos Sampling approach for approximating significance values.

- For size = 3 to 900
  - Take 100,000 samples of P_k : $|P_k|$ = size
  - Score each sample and put them into bins according to their scores

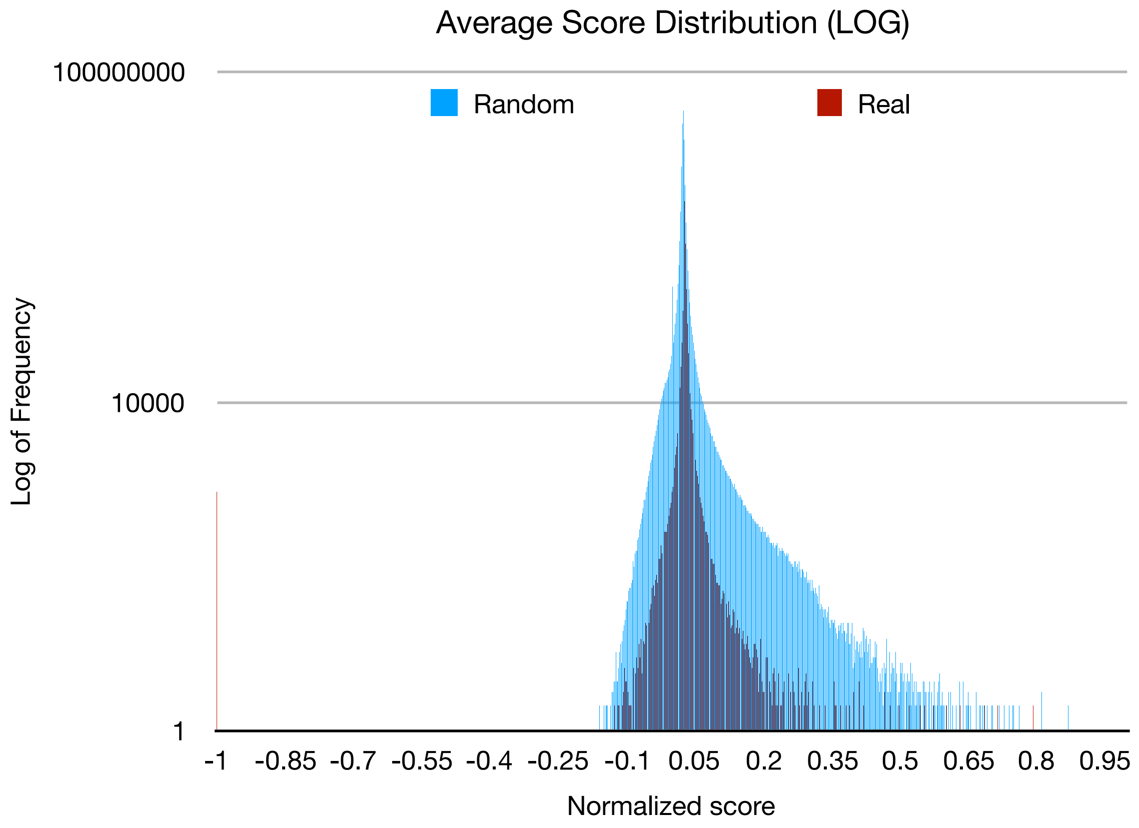
$$p - value(m_i) = \frac{\# P_k : score(P_k) > P_i}{\# P_k}$$

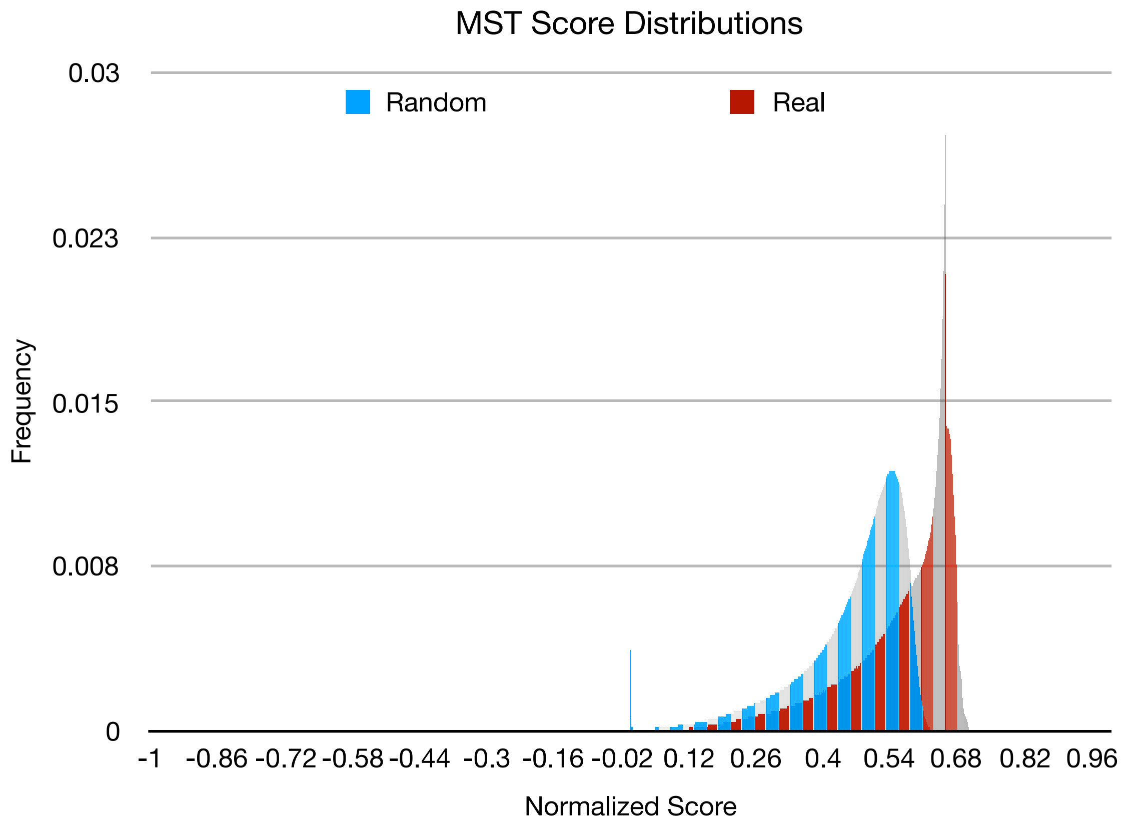
Distribution of scores of samples





Distribution of scores

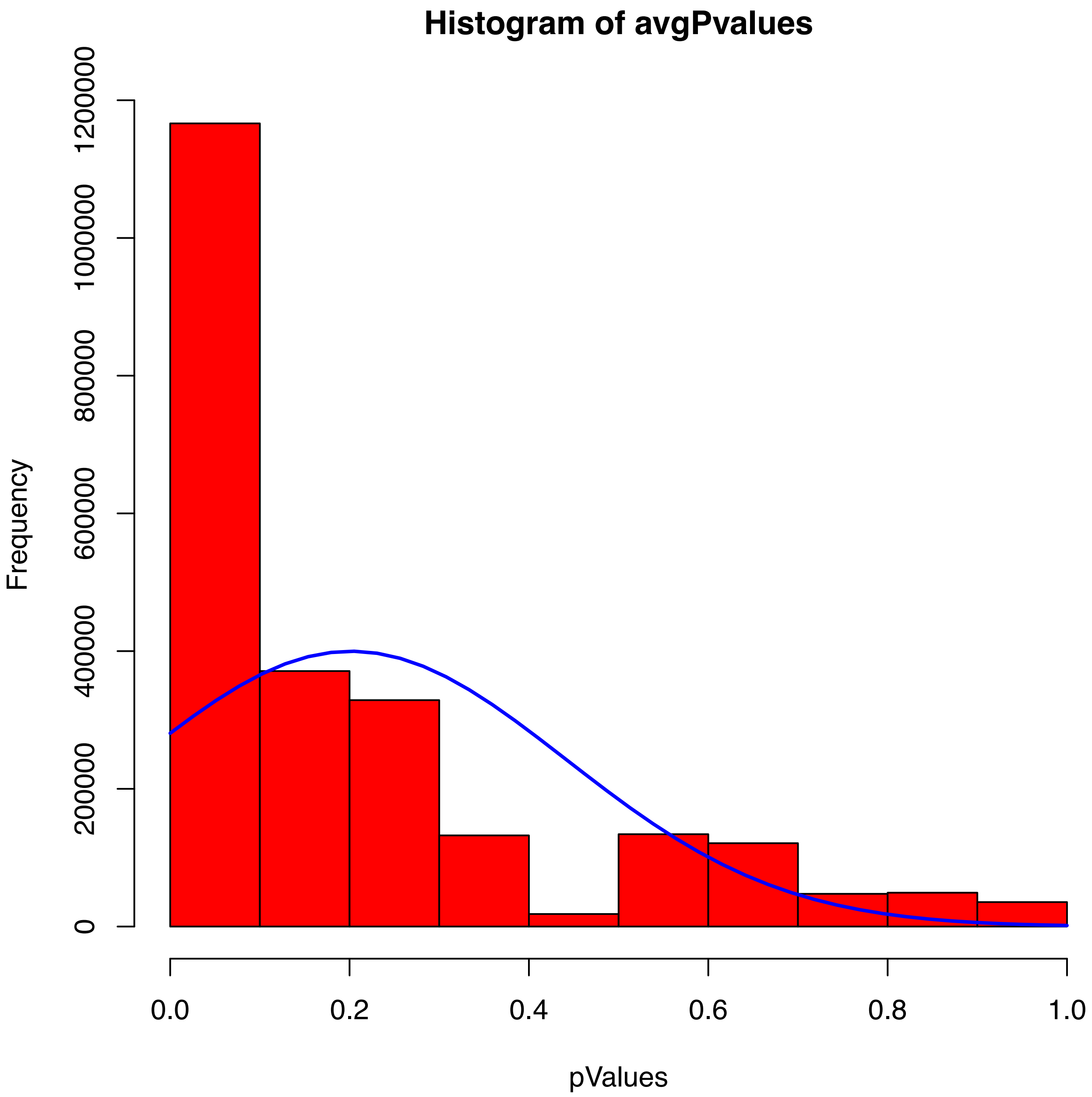




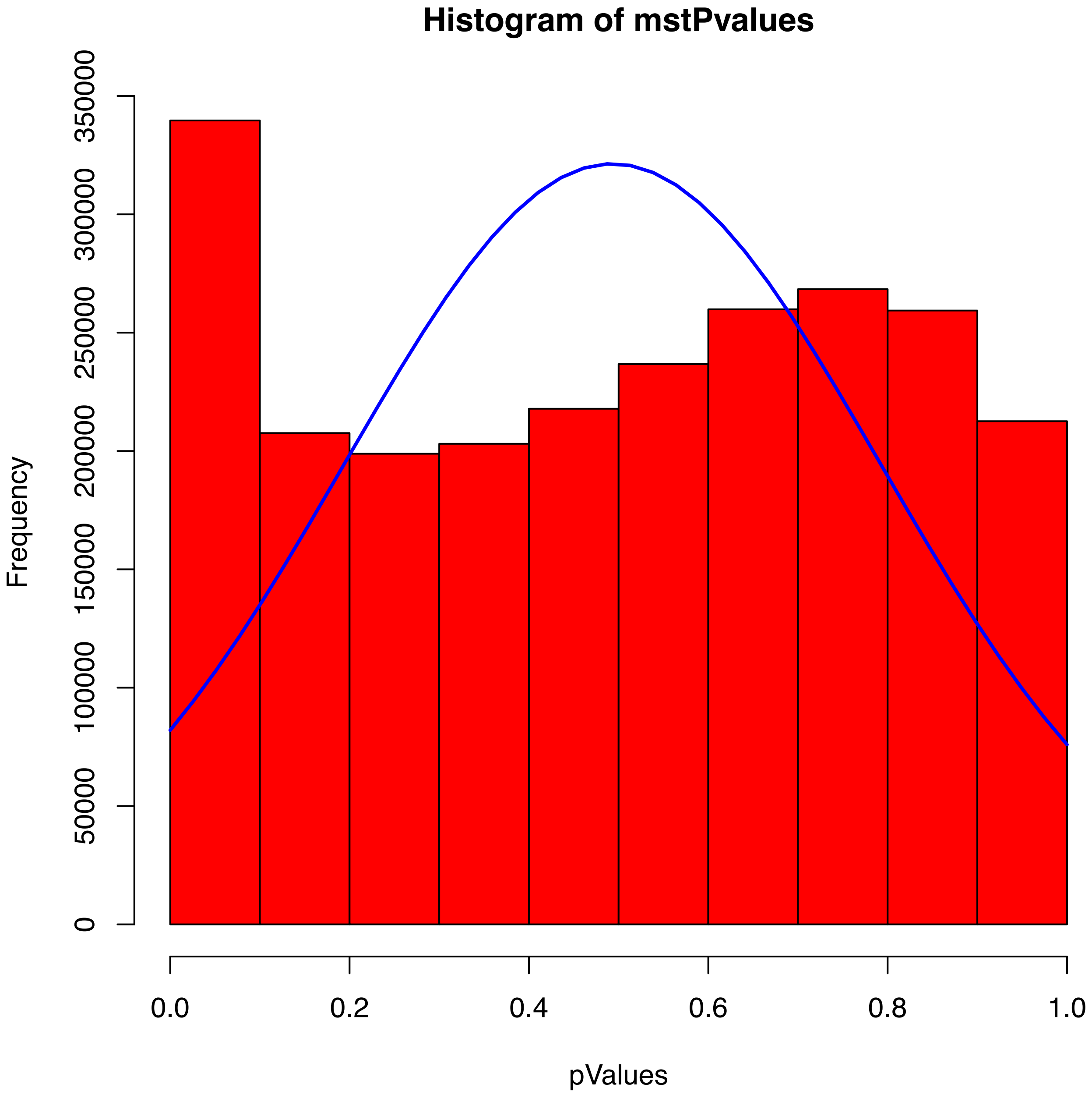
Execution time

- Compute Canada’s Graham (Waterloo) compute cluster
- Scoring the data set:
    - 343 cores * 12h = 4,116 cpu*h
- Scoring the shuffled data set
    - 343 cores * 12h = 4,116 cpu*h
- Scoring the samples
    - 897 cores * 24h = 21,528 cpu*h
- All together
    - 3.4 cpu*years

Results


Histogram of avgPvalues
Frequency
0
200000
400000
600000
800000
1000000
1200000
0.0
0.2
0.4
0.6
0.8
1.0
pValues

Result


Histogram of mstPvalues
Frequency
0
50000
100000
150000
200000
250000
300000
350000
0.0
0.2
0.4
0.6
0.8
1.0
pValues

Correction for Multiple Hypothesis Testing

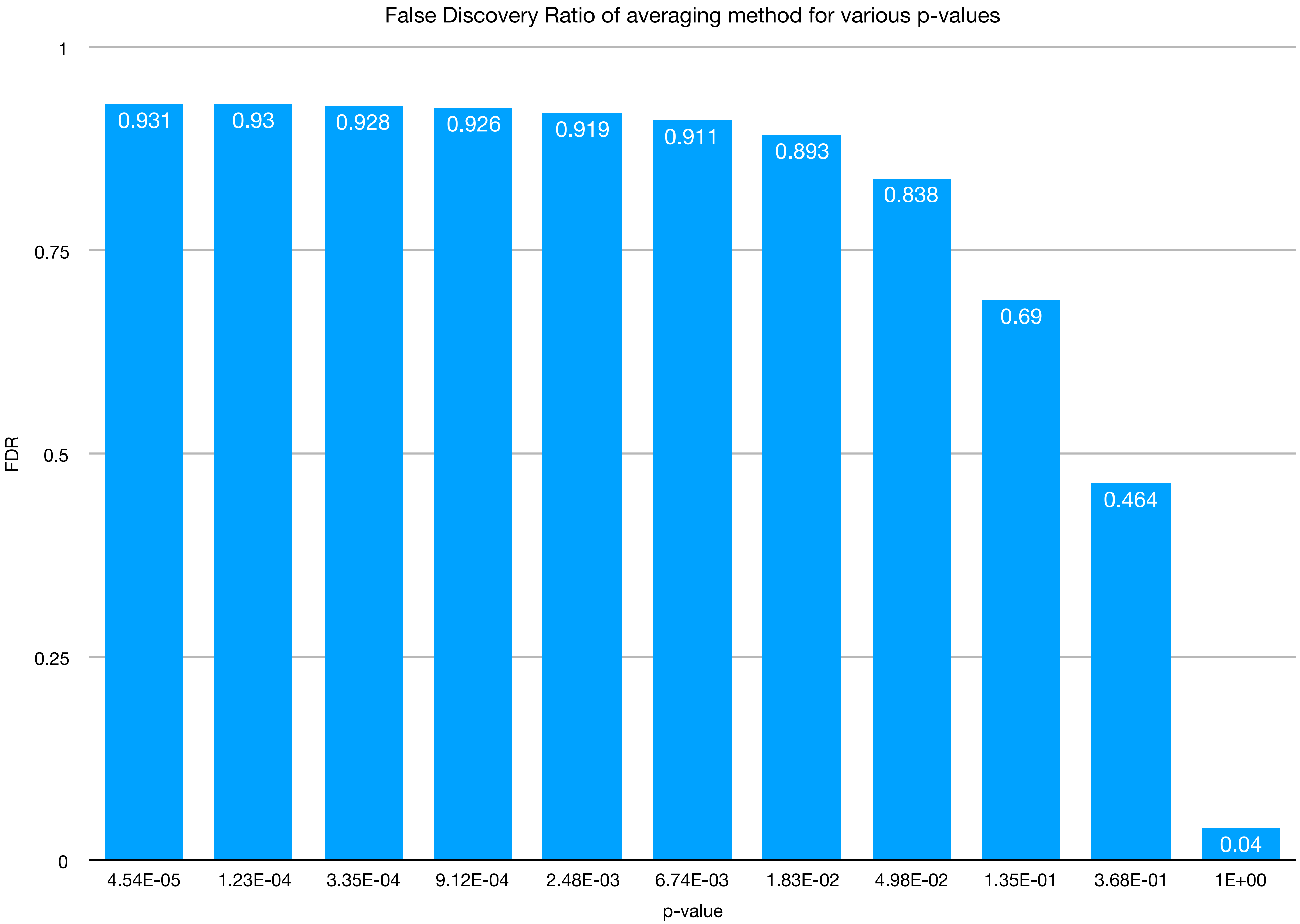
- 2,403,901 highly dependent p-values

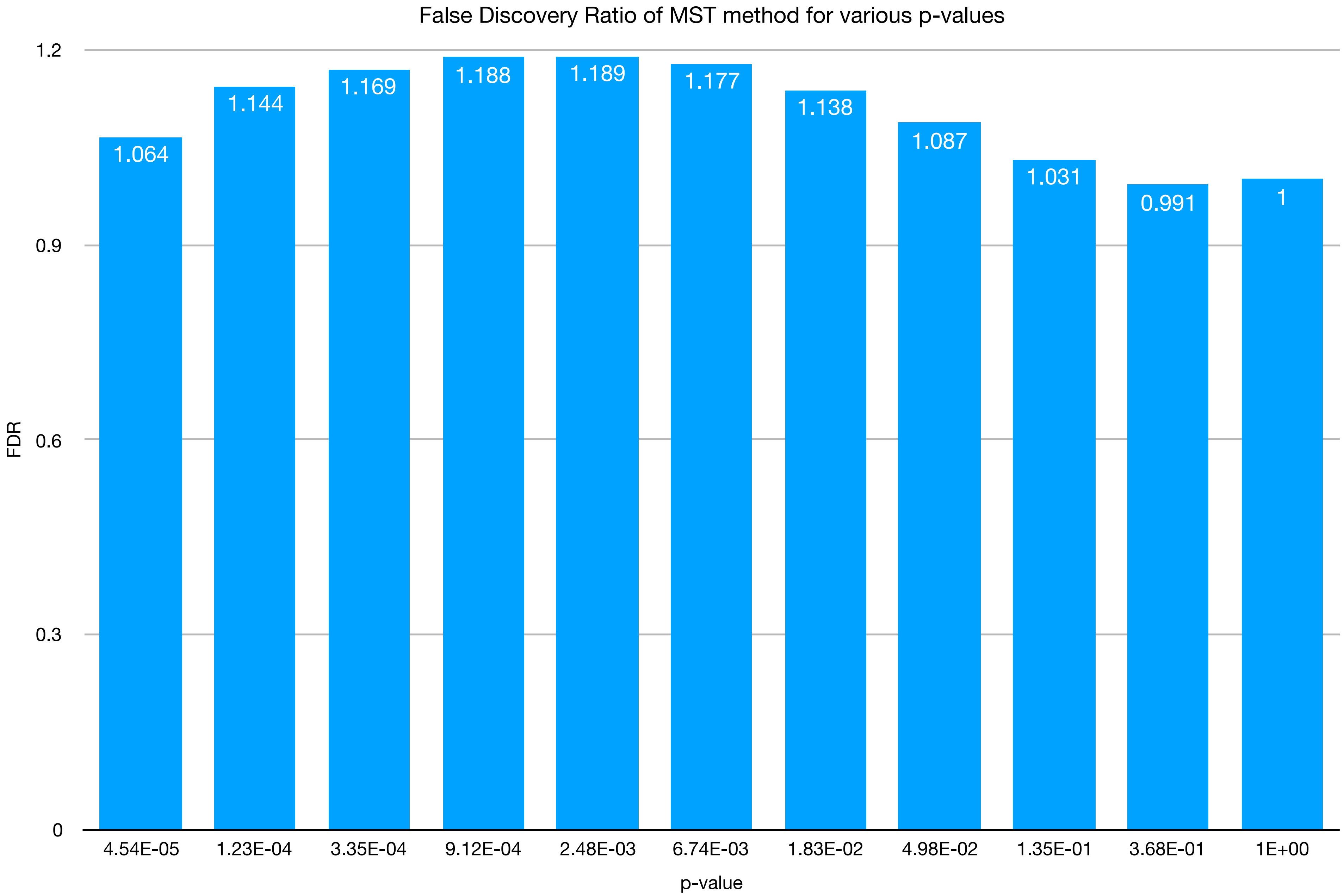
- Need to control of the false discovery rate in multiple testing under dependency

- Benjamini & Yekutieli too strict (no significant values)

Estimating an FDR

- We scramble our dataset and run the algorithms on it.
- For different p-value’s p:
  - Let N(p) denote the number of motifs in the scrambled dataset which have a score at least as significant as p.
  - Let M(p) denote the number of motifs in the original dataset which have a score at least as significant as p.
  - $FDR(p) = \frac{N(p)}{M(p)}$


False Discovery Ratio of averaging method for various p-values
FDR
1
0.75
0.5
0.25
0
0.931
0.93
0.928
0.926
0.919
0.911
0.893
0.838
0.69
0.464
0.04
4.54E-05
1.23E-04
3.35E-04
9.12E-04
2.48E-03
6.74E-03
1.83E-02
4.98E-02
1.35E-01
3.68E-01
1E+00
p-value


False Discovery Ratio of MST method for various p-values
FDR
1.2
0.9
0.6
0.3
0
1.064
1.144
1.169
1.188
1.189
1.177
1.138
1.087
1.031
0.991
1
4.54E-05
1.23E-04
3.35E-04
9.12E-04
2.48E-03
6.74E-03
1.83E-02
4.98E-02
1.35E-01
3.68E-01
1E+00
p-value

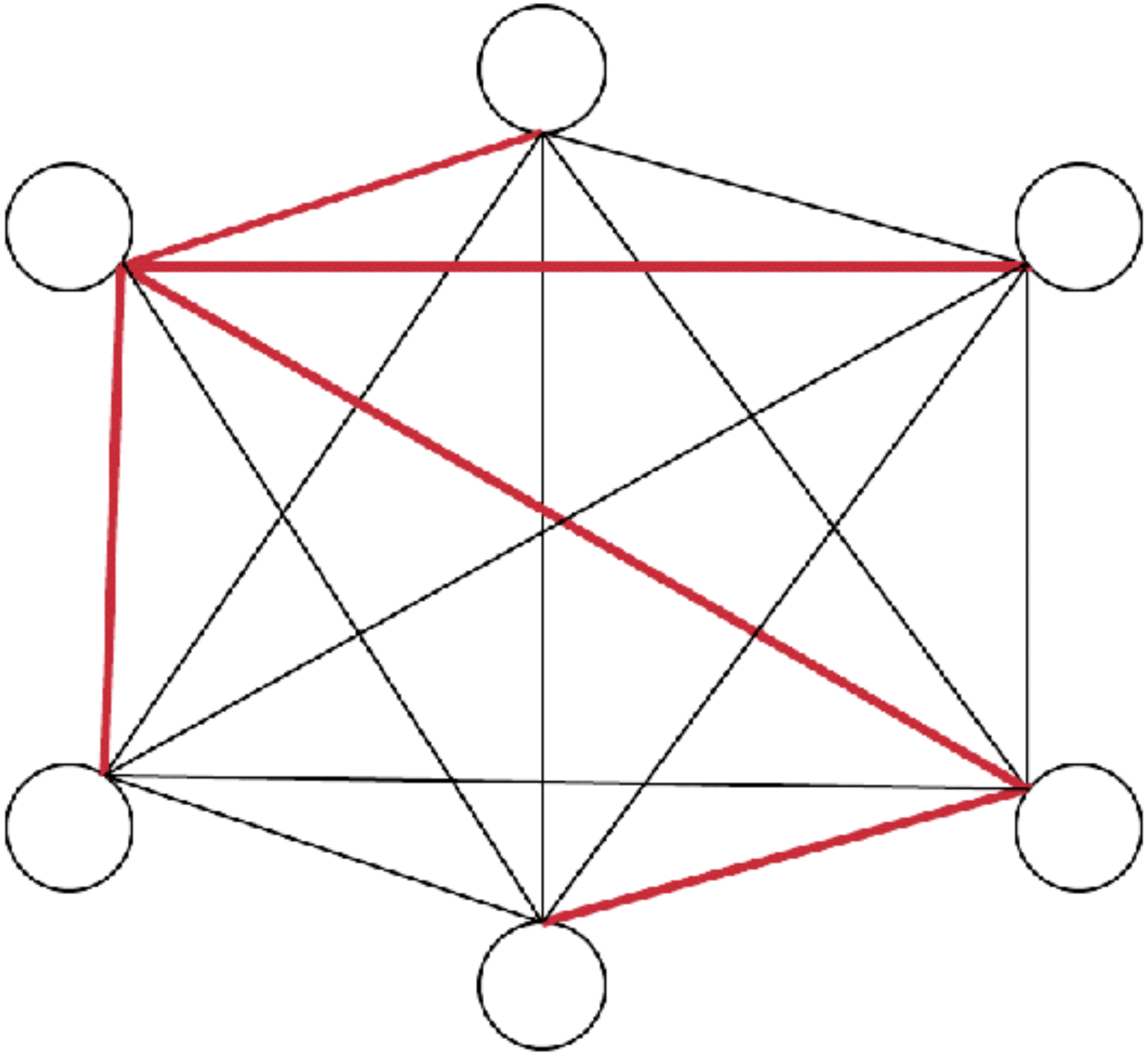
Possible explanations

- Investigating weird behaviours in code
- Shuffling genes changes data characteristics
- Still too much noise
- MST sensitive to outliers (proteins very correlated with every other protein)

Possible improvements

- An algorithm such as Page Rank could reduce noise caused by outlier proteins by reducing their contribution.
- Taking a top fraction of MST edges could further reduce the noise cause by uncorrelated proteins.
- Sample proteins of all sizes to get a more complete picture (instead of 3 to 900)

Outliers



Acknowledgments

- Daniela Sosa, Jack Ryan, Amir Kalani, Linh Nguyen and everyone else at the lab for their help.
- Dr. Mathieu Lavallée-Adam and Dr. Marcel Turcotte
- School of Information Technology and Computer Science (SITE)
- Department of Biochemistry, Microbiology and Immunology (BMI)
- NSERC and Compute Canada

uOttawa
Faculté de médecine
Faculty of Medicine

uOttawa
Faculté de génie
Faculty of Engineering

NSERC
CRSNG

compute
canada