A NETWORK MODEL TO INVESTIGATE ROBUSTNESS OF GENE EXPRESSIONS

Naresh Pasupuleti, Kathryn Cooper
College of IS&T, University of Nebraska at Omaha, Omaha, NE 68182

Objective

- Provide an ideal software pipeline which can supply valid, reproducible and reliable correlation networks.

Motivation and background

- Correlation networks are ideal to describe the relationship between the expression profiles of genes.
- When genes corresponding to a particular part of the body becomes not expressed, it leads to impairment or mutation.
- We want to identify genes that are expressed in concert to be able to identify defective cellular programs.
- By understanding this co-regulation, different ways for the healthy development of a cell can be identified and even changes leading to disease can be detected.
- However, this concept is not fully applied due to lack of benchmarking studies confirming universal applicability.
- Our aim is to explore the robustness of gene expressions by comparing structural similarities of commonly developed networks using big data infrastructures.

Methodology

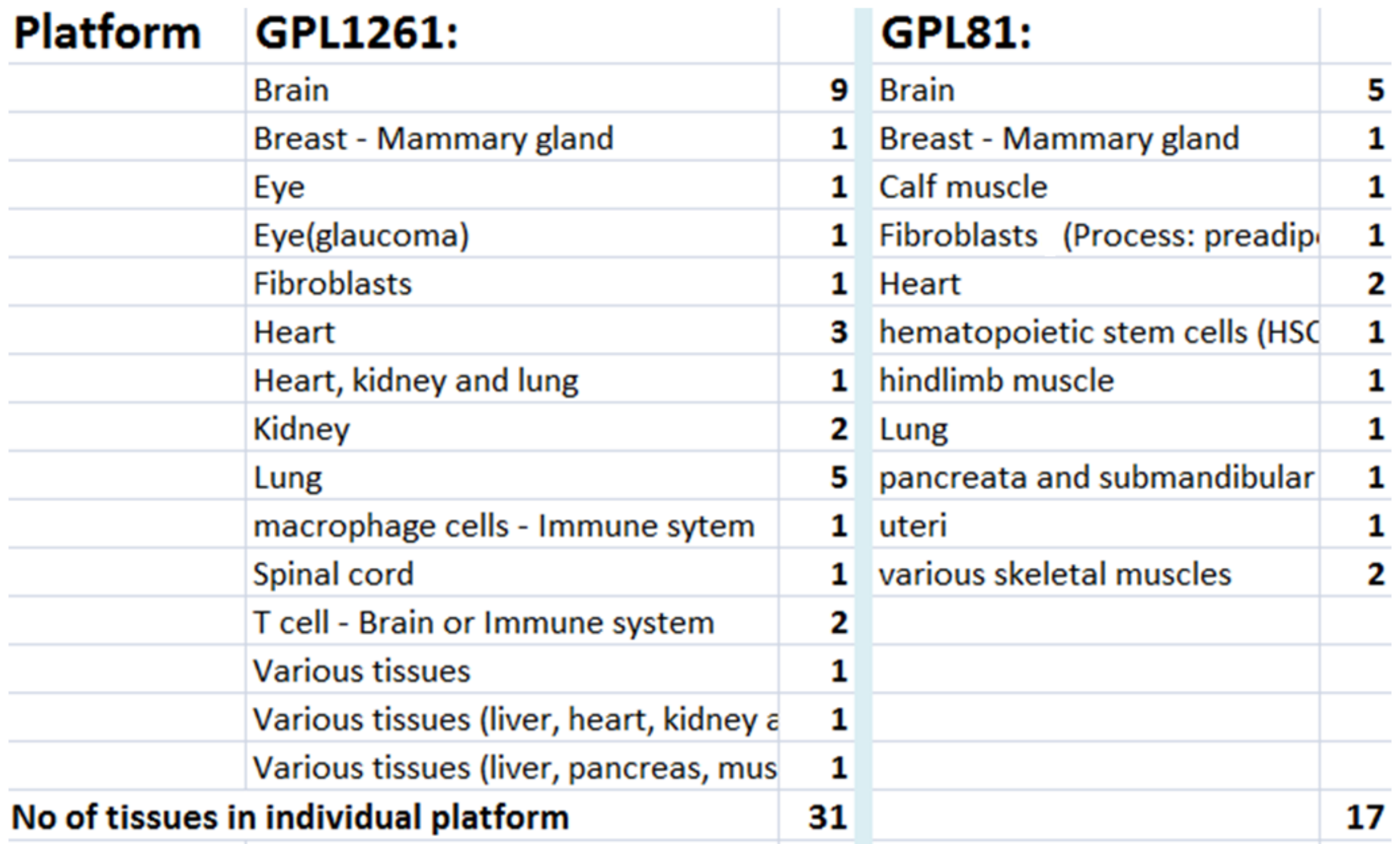
1. Dataset extraction from NCBI
2. Data cleaning and processing
3. Data conversion
4. Finding the overlapping edges
5. Correlation networks creation
6. Statistical data analysis

Hypothesis

- We want to compare genes expressed under different conditions but originating from the same platform and the same organism (*Mus musculus,* or mouse), to see if the correlation is consistent between experiments (because it is always there) or not (because it is unique). Our null hypothesis is that the correlation does not vary widely between similarly constructed datasets.

Dataset Selection

- Studied all the model organisms and selected mouse for our study.
- Used selenium web automation tool to extract top hundred datasets with highest number of samples.
- Finalized GSE 6514 dataset as it is satisfying our criteria of 'same organism (mouse), same platform (GPL1261), same tissue (Brain)'. It has the highest number of samples (90) out of all eligible datasets.

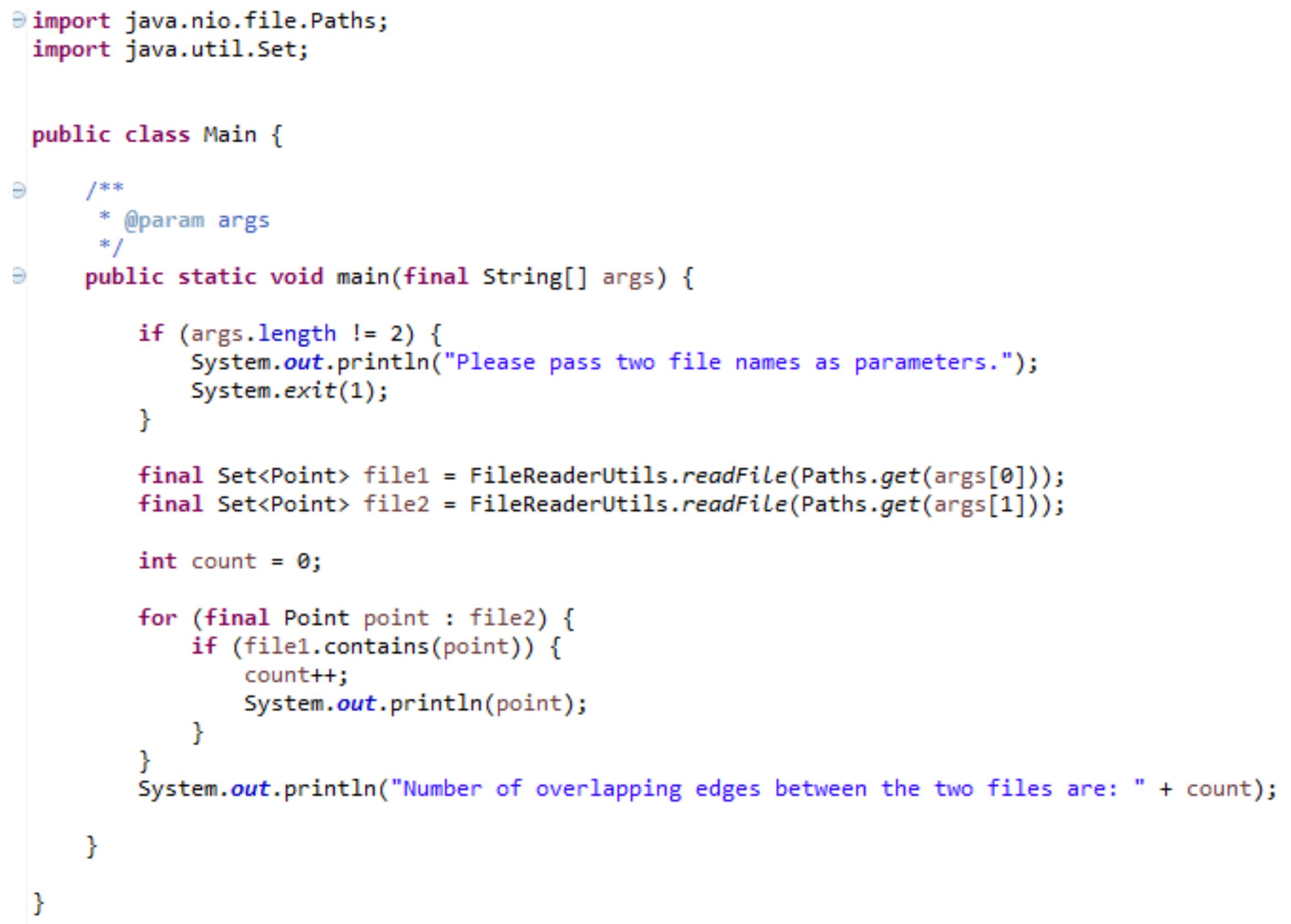


Platform	GPL1261:		GPL81:	
	Brain	9	Brain	5
	Breast - Mammary gland	1	Breast - Mammary gland	1
	Eye	1	Calf muscle	1
	Eye(glaucoma)	1	Fibroblasts (Process: preadip	1
	Fibroblasts	1	Heart	2
	Heart	3	hematopoietic stem cells (HSC	1
	Heart, kidney and lung	1	hindlimb muscle	1
	Kidney	2	Lung	1
	Lung	5	pancreata and submandibular	1
	macrophage cells - Immune sytem	1	uteri	1
	Spinal cord	1	various skeletal muscles	2
	T cell - Brain or Immune system	2		
	Various tissues	1		
	Various tissues (liver, heart, kidney a	1		
	Various tissues (liver, pancreas, mus	1		
No of tissues in individual platform		31		17

Figure 1: Datasets classifications

Data Cleaning and Conversion

- Cleaned the raw dataset using excel and converted the data by suing Perl software from Dr. Cooper to create correlation.
- Subdivided the dataset into 18 small datasets (5 samples each for a dataset)
- Used java program to find out all he possible overlapping edges by pair-wise comparison of all 18 networks, or a total of 153 combinations.



```java
import java.nio.file.Paths;
import java.util.Set;


public class Main {

    /**
     * @param args
     */
    public static void main(final String[] args) {

        if (args.length != 2) {
            System.out.println("Please pass two file names as parameters.");
            System.exit(1);
        }

        final Set<Point> file1 = FileReaderUtils.readFile(Paths.get(args[0]));
        final Set<Point> file2 = FileReaderUtils.readFile(Paths.get(args[1]));

        int count = 0;

        for (final Point point : file2) {
            if (file1.contains(point)) {
                count++;
                System.out.println(point);
            }
        }
        System.out.println("Number of overlapping edges between the two files are: " + count);

    }

}
```

Figure2: Identifying overlapping edges

Use of Neo4j

- Used a tool named Neo4j to create correlation networks which takes edge files as input.
- It is a graph DBMS and a tool to visualize networks.
- Uses Cypher Query Language.

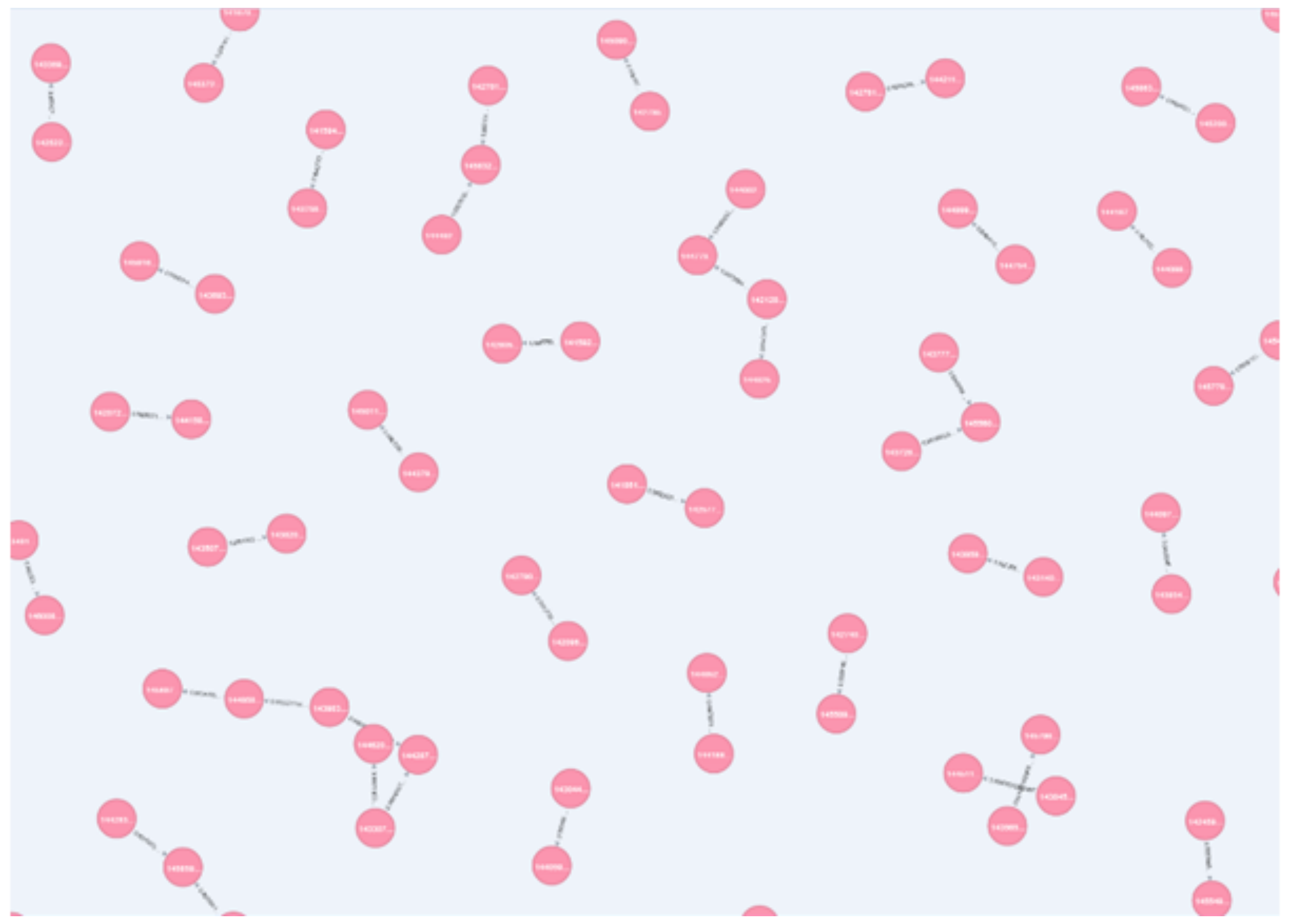


Figure 3: Correlation network image

Results - Networks

- Converted huge network images into tables by using Neo4j
- Analysing all the 153 network images and related tables.
- As of now, we couldn't find any common edge in all of those network files.
- Working on a java program which takes a file with all the unique edges from all of the 153 network tables and gives an output file containing a table with each edge and the name of the network where the edge is from.
- By completing this statistical data analysis, we can form a theory about structure of the correlation networks which supports their conceptual usability in biomedical big data.

References

Dempsey, K., Ali, H. (2011). Evaluation of essential genes in correlation networks using measures of centrality. *Proceedings of the IEEE Bioinformatics and Biomedicine Workshops (BIBMW)*, 509-515.

Dempsey, K., Ali, H. (2012). On the Discovery of Cellular Subsystems in Correlation Networks using Centrality Measures. *Current Bioinformatics* 8(3):305-314.

Dempsey, K., Thapa, I., Bastola, D., Ali, H. (2012). Functional Identification in Correlation Networks using Gene Ontology Edge Annotation. *International Journal of Computational Biology and Drug Design (IJCBDD)*, 5(3-4):222-44.

Newman, MEJ. (2002). Assortative Mixing in Networks. *Phys. Rev. Let.*, 89(20):208701.

Goltsev, V., Dorogovtsev, SN, Mendes, JFF. (2008). Percolation on Correlated Networks. *Physical Review E*, 78(5): 051105.

Kitano, H. (2004). Biological Robustness. *Nat. Rev. Genet.*, 5: 826–837.

"R: Network Analysis and Visualization." In *The iGraph Library* online. Retrieved from http://igraph.sourceforge.net/.

Dempsey, K., Thapa, I., Cortes, C., Eriksen, Z., Bastola, D., Ali, H. On mining biological signals using correlation networks. *IEEE International Conference on Data Mining (ICDM 2014).* December 7-10, 2013; Dallas, TX (Publication pending).