UNIVERSIDADE FEDERAL DO AMAZONAS
INSTITUTO DE COMPUTAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

NETWORK SCIENCE APPROACH FOR ENRICHMENT ANALYSIS IN BREAST AND OVARIAN CANCER

Ficha Catalográfica

Ficha catalográfica elaborada automaticamente de acordo com os dados fornecidos pelo(a) autor(a).

Okimoto, Leandro Youiti Silva
O41n Network Science Approach for Enrichment Analysis in Breast and Ovarian Cancer / Leandro Youiti Silva Okimoto. 2019
62 f.: il. color; 31 cm.

Orientador: Eduardo Freire Nakamura
Orientador: Fabíola Guerra Nakamura
Dissertação (Mestrado em Informática) - Universidade Federal do Amazonas.

1. cancer. 2. breast cancer. 3. ovarian cancer. 4. molecular biology. 5. network analysis. I. Nakamura, Eduardo Freire II. Universidade Federal do Amazonas III. Título

FOLHA DE APROVAÇÃO

"Network Science Approach for Enrichment Analysis in Breast and Ovarian Cancer"

LEANDRO YOUITI SILVA OKIMOTO

Dissertação de Mestrado defendida e aprovada pela banca examinadora constituída pelos Professores:

Prof. Eduardo Freire Nakamura - PRESIDENTE

Prof. Altigran Soares da Silva - MEMBRO INTERNO

Prof. Claudio T. Silva - MEMBRO EXTERNO

Prof. David Fenyo - MEMBRO EXTERNO

Manaus, 04 de Junho de 2019

LEANDRO YOUITI SILVA OKIMOTO

NETWORK SCIENCE APPROACH FOR ENRICHMENT ANALYSIS IN BREAST AND OVARIAN CANCER

Dissertação apresentada ao Programa de Pós-Graduação em Informática do Instituto de Computação da Universidade Federal do Amazonas, Campus Universitário Senador Arthur Virgílio Filho, como requisito parcial para a obtenção do grau de Mestre em Informática.

Orientador: Eduardo Freire Nakamura
Co-orientador: Fabíola Guerra Nakamura

Manaus - AM
Abril de 2019

LEANDRO YOUITI SILVA OKIMOTO

NETWORK SCIENCE APPROACH FOR ENRICHMENT ANALYSIS IN BREAST AND OVARIAN CANCER

Dissertation presented to the Graduate Program in Informatics of the Federal University of Amazonas in partial fulfillment of the requirements for the degree of Master in Informatics.

Advisor: Eduardo Freire Nakamura
Co-orientador: Fabíola Guerra Nakamura

Manaus - AM
April 2019

"Science and everyday life cannot and should not be separated."
(Rosalind Franklin)

Abstract

The imprecise identification of cancer characteristics can lead patient to aggressive and unnecessary treatments. Therefore, it is crucial to identify tumor intrinsic characteristics more precisely to propose individual-tailored treatment. In this work, we present a brief explanation of fundamentals and researches in computer graph theory that seek to solve problems of identification, classification, and characterization of certain cancer types. We proposed a novel solution based on Network Science to find list of genes for enrichment analysis in Breast and Ovarian cancer using proteogenomic information. In our results, we show that our approach is capable of capturing biological processes and sets of genes related to cancer and other processes, which opens a range of possibilities for further studies.

Keywords: cancer, breast cancer, ovarian cancer, molecular biology, network analysis.

List of Figures

List of Tables

Contents

1

Introduction

Breast cancer is a heterogeneous disease with subtypes that present distinct biological characteristics. These differences affect the response to treatment and can lead to different clinical outcomes [Yersal and Barutca, 2014]. One important task is to determine the subtype of breast cancer to choose the right treatment. Not rarely, one can find cases where low-risk patients receive aggressive or unresponsive treatment.

Ovarian cancer is the most lethal malignancy of the female reproductive system. It causes over 14,000 deaths in the US and 114,000 worldwide, annually [Pecorelli et al., 2003]. Standard therapy results in complete response for 70% of patients. However, most will relapse within 18 months because of chemoresistant disease [Ozols, 2005]. Thus, one task is to improve targeted therapies strategies to reduce mortality rates.

Perou et al. [1999] analyze breast tumor tissues to find patterns of gene expression through gene expression micro-arrays, with good accuracy. Nowadays, his work is used as consensus for Breast Cancer subtypes.

Mertins et al. [2016], Ozols [2005], Yersal and Barutca [2014] declare tumors with similar clinical and pathological presentations may have different behavior. So, cancer treatment should be specific and individual to each patient. Analyses of breast cancer with new molecular techniques now hold promise for the development of more accurate tests to predict recurrence [Yersal and Barutca, 2014].

The aim of this research is to propose and evaluate topological properties of Network Science models for Enrichment Analysis in Breast and Ovarian cancer networks built from gene expression data. We can use Enrichment Analysis to find list of genes in a collection of annotated gene sets, related or not to cancer.

1.1 Motivation

Breast cancer subtypes lead to differences in patterns of response to various treatment modalities [Yersal and Barutca, 2014]. The prediction of these therapies responses using molecular attributes are key to cancer biology [Liu et al., 2014]. In this context, several papers analyze efficient ways to characterize breast and ovarian cancer.

The 12° St. Gallen International Breast Cancer Conference (2011) classifies patients for therapeutic purposes based on the recognition of subtypes of Breast cancer per spectrum [Goldhirsch et al., 2011]. Recommendations for systemic therapy follow the subtype classification proposed by Perou et al. [2000]. This classification uses biological markers or biomarkers in short, which are shown in Table 1.1, where the expressions of each biomarker classify the tumors into subtypes. After this conference, the experts reached a consensus on using this approach for primary treatment of Breast Cancer.

Subtype *IHC*	**Biomarkers**	**Therapy**
Luminal A	HR+/HER2-/Ki67low	*Endocrine therapy*
Luminal B	HR+/HER2-/Ki67high	*Endocrine therapy +- cytotoxic therapy*
	HR+/HER2+	*Cytotoxics therapy + anti-HER2 + hormonal therapy*
HER2-positive	HR-/HER2+	*Cytotoxics + anti-HER2 therapy*
Triple negative (Basal)	HR-/HER2-	*Cytotoxic therapy*

Table 1.1: Classification accordingly to Perou et al. [2000], where the subtypes are based in *immunohistochemistry (IHC).*

Although the consensus was established, current studies show how much further research is needed to achieve an optimal model to classify Breast cancer subtypes [Mertins et al., 2016]. The proposed models still have a significant percentage of error when applied to subtypes that share almost same characteristics but have different prognostic.

For Ovarian cancer, we still need better ways to refine cancer subtype characteristics and treatments. Accordingly to biological analysis, breast cancer is more consolidated in subtype manners. In our research, we use a gene expression dataset containing breast cancer with consolidated subtypes in literature and ovarian cancer without subtype specifications.

1.2 Problem

Types of cancer lead to differences in patterns of response to various treatment modalities. The prediction of these therapies responses using molecular attributes are the

key to cancer biology. One of the main challenges for those predictions is the "curse of dimensionality", this is when we have too many features for few samples. Because lack of samples, machine learning may not be a good approach.

Although we may not use machine learning due to lack of samples and high number of features (genes and proteins), we can use the features for better characterization. An idea is using enrichment analysis that identify classes of genes or proteins that are over-represented in a large set of genes or proteins. One hypothesis is to extract characteristics from networks, that provides a natural representation of a biological system, and use them as inputs for enrichment analysis.

1.3 Objective

This research aims to propose and evaluate models for Breast and Ovarian cancer characterization applying Enrichment Analysis in WCN (Weighted Correlation Network). We use gene expression data from tumor tissues to build networks. For this purpose, we define specific objectives, which include:

- To show the existence and quantify the statistical correlation between gene expression data;
- To evaluate the existence of patterns between groups of gene expression data for Breast and Ovarian cancer;
- To evaluate communities from Community Detection as input for Enrichment Analysis;
- To evaluate ranked lists of centrality measures as input for Enrichment Analysis.

1.4 Contributions

In this work, we use an approach for enrichment pathways analysis using network science. We compared communities between Breast and Ovarian cancer networks. We evaluated centrality measures as inputs to enrichment analysis.

The contributions of our work are: (i) community detection, applied to WCN, find biologically cohesive group of genes; and (ii) we found that some centrality measures can show cancer related gene sets in their ranked lists.

1.5 Document Outline

In chapter 2 we present the theoretical basis necessary for the understanding of the adopted methods and the related works, showing existing methods and the advance of the most current researches. In Chapter 3 we present the proposed solution where we explain our method. In chapter 4 we present the partial results achieved in this work. Finally, in chapter 5 we discuss our conclusions and some future activities.

2

Fundamentals

In this chapter we present concepts necessary for the work development. We divide the fundamentals in 4 sections: Biological Background, Weighted Correlation Network, Network Analysis, and Related Works.

Section 2.1 shows an overview about Gene Expression Data and Gene Set Enrichment Analysis, they are necessary for our understanding because it comprehends our data format and our biological analysis. Section 2.2 presents the basis of our work as structure, we analyze using properties of Network Analysis in section 2.3. Section 2.4 we present some related works and finally in section 2.5 present some discussion about this chapter.

2.1 Biological Background

One of the major challenges of current cancer biology is the development of personalized diagnostic and therapeutic strategies [Mirnezami et al., 2012]. In the last decades, the increasing availability of data regarding genomic and proteomic profiles of cancer patients has provided a new source of essential information, which however needs efficient theoretical frameworks, instruments, and computational tools to be exploited [Graudenzi et al., 2017].

2.1.1 Gene Expression Data

Gene expression data is the information of the process by which a gene is used in the synthesis of a functional gene product. The advances in omic techniques provide unprecedented capacity to measure gene expression data as RNA, protein, and post-translational modification levels under different biological contexts such as time, cell

states, tissues or organisms [Consortium et al., 2012, Kundaje et al., 2015, Lonsdale et al., 2013, Weinstein et al., 2013]. These precise information are important for our study to reconstruct part of a network based on the central dogma of molecular biology [Crick, 1970], three sets of information: CNA (Copy Number Alteration), RNA (Ribonucleic Acid), and Proteins.

2.1.2 Gene Set

In our work, we define a Gene Set as a sorted list of genes responsible for a specific biological function. For example, it could be cancer genes involved in a specific pathway or specific biological occurrency. Gene sets are lists of genes that were studied in a published work and annotated in a Database [Liberzon et al., 2011].

2.1.3 Enrichment Analysis

Enrichment analysis focus on identifying groups of genes that together act as biological function, chromosomal location, or regulation. This analysis is important to identify group of correlated genes, analysing its biological background [Khatri et al., 2012]. Then, it is possible to embody characteristics related to gene sets that are in cancer-related data sets.

2.2 Weighted Correlation Network

A long-standing problem in biological systems is to infer causal, regulatory connections among genes, proteins, and metabolines [Chasman et al., 2016]. A network provides a natural representation of a complex cellular system with nodes representing the molecular components and edges representing different connections. A simple solution to create edges between nodes is by correlating them and use the coefficient values as weight for edges. One of the most commonly used correlation methods for gene expression data is Pearson Correlation Coefficient [D'haeseleer, 2005], the reason is the existence of a linear correlation in gene expression data. WCN is this network built from correlation coefficients as edges and correlated variables as nodes.

Pearson Correlation Coefficient

Pearson Correlation Coefficient is a measure of linear correlation between two variables [Benesty et al., 2009], its properties are described below.

- Measures the linear relationship between X and Y;
- Range: $-1 \leq r \leq 1$;
- Correlation coefficient is a unitless index of strength of association between two variables (+ = positive association, - = negative, 0 = no association).

Pearson Correlation Coefficient formula is:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma_{i=1}^{n}(x_i - \bar{x})^2 \Sigma_{i=1}^{n}(y_i - \bar{y})^2}}, \tag{2.1}$$

in which X and Y are the variables, r is the correlation coefficient of pair X and Y, and n is the sample size.

The reason behind the WCN in cancer is to use network language to describe the relationships between biomolecules. Although there are other statistical techniques for the analysis of correlation matrices, the language of the network is particularly intuitive for biologists and allows simple analyzes of the informative network. We can use those networks to meet many analytical objectives, for example finding clusters of interconnected biomolecules.

2.3 Network Analysis

Our network will have important topological properties to evaluate a vertex and their connections. We call those properties as part of network analysis. In this work, we are going to explore the network topology, communities of genes and important node ranking, to evaluate as input for enrichment analysis.

In the following subsections, we present two concepts that are applied in our Weigthed Correlation Network: Community Detection and Centrality Measures.

2.3.1 Community Detection

The process of discovering cohesive groups or clusters in a network is known as community detection. Detecting communities in biological networks can be useful in applications where the specific characteristics of each group are evaluated, for example, recommending a certain biological process expressed in a community [Atay et al., 2017].

Community Detection based in Clustering

The main objective of this approach is to detect clusters, cohesive groups or subgroups. Among the innovators of community detection in this approach are Girvan and Newman [2002], they proposed an algorithm using *betweenness centrality* (Explanation in item 5 of this work), in which vertexes with a greater value of this measure are used as cutting vertexes for the construction of related components. Another concept defined by Girvan and Newman [2002] was the measure known as 'modularity' to quantify the quality of communities [Newman and Girvan, 2004]. Blondel et al. [2008] inspired by modularity, created the 'modularity of louvain', which is an optimized method for community detection. Modularity of louvain is a greedy heuristic algorithm that builts communities and uses modularity as a metric to evaluate quality of the communities.

2.3.2 Centrality Measures

In order to identify the importance of entities, complex networks use a concept of centrality measure, which are measures developed to determine the importance of vertexes and edges, using structural characteristics of networks. The following items in this subsection list some methods and present some works that used it.

Weighted Degree Centrality

This measure is similar to the Degree Centrality measure, the difference being that this measure takes into account the weight assigned to the edges. In the paper by Tang et al. [2014], the authors show that it is possible to predict essential proteins using Weighted Degree Centrality, the authors used Pearson correlation coefficient (PCC) for the computation of the edge weights. As shown in Algorithm 1, the time complexity of this algorithm is $O(E)$ in which E is the number of edges.

Algorithm 1: Weighted_Degree(G,n)

```
1 begin
      // will compute E times → O(E)
2     for each e ∈ G[E] do
3         e.from ← e.from + weight(e)
4         e.to ← e.to + weight(e)
5     end
6 end
```

Eigenvector Centrality

This measure is similar to the measure Weighted Degree Centrality with a return: a vertex receives the sum of the importance values of its neighbors. Then the importance of a vertex is the weighted sum of the importance of the neighboring vertexes added to its value of importance. The importance x_i of a vertex i receives the weighted sum of the importance of its neighbors: $x_i = \sum_{j \in V} w_{ji} x_j$ for each $i \in V$. In the work of Borgatti [1995], the author explains that by using the *Eigenvector Centrality* measure we are expanding the understanding of the risk of infection, since a person A that has several people in the network has a great chance of infecting everyone else who has a relationship, even if that relationship is only A. The time complexity of this algorithm (shown in Algorithm 2) is $O(E)$, in which E is the number of edges.

Algorithm 2: Eigenvector(G,n)

```
1  begin
       // will compute 2 × E times → O(E)
2      for each e ∈ G[E] do
3          V(e.from) ← V(e.from) + weight(e)
4          V(e.to) ← V(e.to) + weight(e)
5      end
6      for each e ∈ G[E] do
7          V(e.from) ← V(e.from) + V(e.to)
8          V(e.to) ← V(e.to) + V(e.from)
9      end
10 end
```

Closeness Centrality

It is a measure that computes the average distance from one vertex to all others, the smaller this distance, the more important is the vertex. In the work of Okamoto et al. [2008], the authors present a comparison of Closeness centrality with PageRank centrality [Beveridge and Shan, 2016]. For the ranking systems, they show that the Closeness centrality has a low computational expense compared to *PageRank* since the measure needs to calculate the *PageRank* value of all vertexes, even if only a sample k of elements is desired. The time complexity of this algorithm (shown in Algorithm 3) is $O(V^2E)$, in which E is the number of edges and V is the number of vertices.

Algorithm 3: Closeness(G,n)

```
1 begin
      // compute V times
2     for each v ∈ G[V] do
          // compute all short paths from v to all other nodes → VE
3         v ← shortest_path(v, G) ÷ size(G[V])
4     end
      // will compute V × V × E → O(V²E)
5 end
```

Betweenness Centrality

The idea of this centrality is to evaluate the importance of a node according to the interposition of a vertex and the path between the other vertexes on the network. Then, the importance of the node is due to the number of paths between two vertexes that cross the chosen vertex. This measure calculates the frequency of trips that pass through that vertex. The betweenness z_i of a vertex i is given by

$$z_i = \sum_{j,k \in V} \frac{\sigma_{jk}(i)}{\sigma_{jk}}, \tag{2.2}$$

in which σ_{jk} is the number of *(j,k)* - smaller paths and $\sigma_{jk}(i)$ the number of these smaller paths that pass through the vertex i [Beveridge and Shan, 2016]. The work of Leydesdorff [2007] has an interesting application using this measure, the author uses *Betweenness centrality* as an indicator of interdisciplinary of scientific journals. The time complexity of this algorithm (shown in Algorithm 4) is $O(V^2E)$, in which E is the number of edges and V is the number of vertices.

Algorithm 4: Betweenness(G,n)

```
1 begin
      // compute V times
2     for each v ∈ G[V] do
          // compute all short paths from v to all other nodes → VE
3         shortest ← shortest_path(v, G)
4     end
5     bet(G[V]) ← frequency(shortest, G[V])
      // compute frequency of shortest paths through each node
      // will compute V × V × E → O(V²E)
6 end
```

Clustering Rank

Mathematically, the Clustering Rank score s_i of node i is defined as:

$$s_i = f(c_i) \sum_{j \in \tau_i} (k^j_{out} + 1), \tag{2.3}$$

in which the term $f(c_i)$ accounts for the effect of i's local clustering and the term "+ 1" results from the contribution of j itself. Here $f(c_i) = 10^{-c_i}$ [Chen et al., 2013]. This centrality is a local ranking algorithm, which takes into account not only the number of neighbors and the neighbors' influences, but also the clustering coefficient. The time complexity of this algorithm (shown in Algorithm 5) is $O(V^2)$, in which E is the number of edges and V is the number of vertices.

Algorithm 5: Clustering(G,n)

```
1 begin
    // compute V times
2   for each v ∈ G[V] do
      // compute cluster rank from node to neighbors → V
3     for each nei ∈ neighborhood(v) do
4       cluster_rank(v) ← cluster_rank(v) + degree(v, nei, loops(v, nei))
5     end
6   end
    // will compute V × V → O(V^2)
7 end
```

Diffusion Degree

Diffusion degree C_{DD} of node v is defined as:

$$C_{DD}(v) = \lambda_v \times C_D(v) + \sum_{i \in neighbors(v)} \lambda_i \times C_D(i), \tag{2.4}$$

in which C_D is degree of of vertex and λ is propagation probability of vertex. In a diffusion process, a node v with propagation probability λ_v, can activate its neighbor u with probability λ_v. When the diffusion process propagates to the next level, active neighbors of v will try to activate their inactive neighbors. Thus the cumulative contribution in the diffusion process by neighbors of v will be maximized when all of its neighbors will be activated in the previous step [Pal et al., 2014]. As shown in Algorithm 6, the time complexity of this algorithm is $O(V^2)$ in which V is the number of nodes.

Algorithm 6: Diffusion(G,n)

```
1 begin
      // apply probability λ for each E
2     d ← degree(G[E]) × λ
      // compute V times
3     for each v ∈ G[V] do
          // compute diffusion from node to neighbors → V
4         for each nei ∈ neighborhood(v) do
5             dif(v) ← dif(v) + d(v, nei)
6         end
7     end
      // will compute E + V × V → O(V^2)
8 end
```

DMNC

One major task in the post-genome era is to reconstruct proteomic and genomic interacting networks using high-throughput experiment data. To identify essential nodes/hubs in these networks is a way to decipher the critical keys inside biochemical pathways or complex networks. These essential nodes/hubs may serve as potential drug-targets for developing novel therapy of human diseases, such as cancer or infectious disease caused by emerging pathogens [Lin et al., 2008].

The Density of Maximum Neighborhood Component (DMNC) was developed for exploring and identifying hubs/essential nodes from networks. The score of node v using $DMNC(v)$, is defined to be $\frac{E}{N^{\epsilon}}$:

$$\frac{|E(MNC(v))|}{|V(MNC(v))|^{\epsilon}}, \tag{2.5}$$

in which for some $1 \leq \epsilon \leq 2$, ϵ is set to be a value of neighborhood control, E is the number of edges, V the number of vertices. MNC is the Maximum Neighborhood Component in which the score of node v, MNC(v), is defined to be the size of the maximum connected component of N(v), the neighborhood N(v) is the set of nodes adjacent to v and does not contain node v. The time complexity of this algorithm (shown in Algorithm 7) is $O(V^2)$, in which V is the number of vertices.

Algorithm 7: DMNC(G,epsilon)

```
1 begin
      // clusters function costs O(V^2)
2     c ← clusters(G)
      // compute V times
3     for each v ∈ G[V] do
4         sub ← subgraph_neighborhood(v)
5         ec ← edge_count(sub, max(c))
6         dmnc(v) ← ec ÷ max(c)^epsilon
7     end
      // will compute V^2 + V → O(V^2)
8 end
```

Laplacian Centrality

The Laplacian centrality with respect to v is:

$$C_v^L = (\Delta E)_v = d_G^2(v) + d_G(v) + 2 \sum_{v_i \in N(v)} d_G(v_i), \tag{2.6}$$

in which G is a graph of n vertices, $N(v)$ is the set of neighbors of v in G and $d_G(v_i)$ is the degree of v_i in G. Laplacian centrality is a simple centrality measure that can be calculated in linear time. It is defined as the drop in the Laplacian energy (i.e. sum of squares of the eigenvalues in the Laplacian matrix) of the graph when the vertex is remove [Qi et al., 2012]. As shown in Algorithm 8, the time complexity of this algorithm is $O(V^2)$ in which V is the number of nodes.

Algorithm 8: Laplacian(G,n)

```
1 begin
      // compute V times
2     for each v ∈ G[V] do
          // sum of neighbors degree V times (worst case)
3         for each nei ∈ neighborhood(v) do
4             deg_nei ← deg_nei + degree(nei)
5         end
6         deg ← degree(v)
7         laplacian(v) ← deg^2 + deg + 2 × deg_nei
8     end
      // will compute V × V → O(V^2)
9 end
```

Leverage

Leverage centrality considers the degree of a node relative to its neighbors and operates under the principle that a node in a network is central if its immediate neighbors rely on that node for information. Leverage centrality of vertex i defined as:

$$l_i = \frac{1}{k_i} \sum_{N_i} \frac{k_i - k_j}{k_i + k_j}, \tag{2.7}$$

in which k_i is degree of a given node i, k_j is degree of each of its neighbors and N_i is all neighbors. A node with negative leverage centrality is influenced by its neighbors, as the neighbors connect and interact with far more nodes. A node with positive leverage centrality, on the other hand, influences its neighbors since the neighbors tend to have far fewer connections [Joyce et al., 2010]. As shown in Algorithm 9, the time complexity of this algorithm is $O(V^2)$ in which V is the number of nodes.

Algorithm 9: Leverage(G,n)

```
1 begin
    // compute V times
2   for each v ∈ G[V] do
      // compute leverage from node to neighbors V times (worst case)
3     for each nei ∈ neighborhood(v) do
4       lev_rank(v) ← lev_rank(v) + (degree(v) − degree(nei)) ÷
          (degree(v) + degree(nei))
5     end
6     lev_rank(v) ← lev_rank(v) ÷ degree(v)
      // will compute V × V → O(V^2)
7   end
8 end
```

Topological Coefficient

The topological coefficient is a relative measure for the extent to which a node shares neighbors with other nodes. Topological coefficient T_n of a node n with k_n neighbors defined as:

$$T_n = \frac{mean(J(n,m))}{k_n}, \tag{2.8}$$

in which $J(n,m)$ is defined for all nodes m that share at least one neighbor with n. The value $J(n,m)$ is the number of neighbors shared between the nodes n and m, plus one if there is a direct link between n and m. Nodes that have one or no neighbors are

assigned a topological coefficient of zero [Assenov et al., 2007]. The time complexity of this algorithm (shown in Algorithm 10) is $O(V^3)$, in which V is the number of vertices.

Algorithm 10: Topological(G,n)

```
1  begin
       // compute V times
2      for each v ∈ G[V] do
3          J ← 0
           // compute node to neighbors V times (worst case)
4          for each nei1 ∈ neighborhood(v) do
               // compute neighbor node to neighbors V times (worst case)
5              for each nei2 ∈ neighborhood(nei1) do
6                  if nei2 ∈ neighborhood(v) then
7                      J ← J + 1
8                  end
9              end
10         end
11         J ← J ÷ shared_neighbors(v)
12         topological(v) ← J ÷ size(neighborhood(v)
           // will compute V × V × V → O(V^3)
13     end
14 end
```

2.4 Related Works

In this section, we show some works about cancer classification and characterization that applied networks on their studies. Those related works are not only for breast and ovarian cancer, it is interesting to analyze works not only applied for only one type of cancer, as we are looking which methods are capable of characterizing general types of cancer.

In Weigthed Correlation Network, usually the authors' strategy is to correlate molecules and find patterns that may be useful in discriminating some characteristics of cancer. They construct the networks and use the topology of these networks to identify discriminative characteristics.

2.4.1 State-of-the-art

Zhang et al. [2016] present a classification procedure for ovarian cancer and identify networks and genes for each subtype by integrating multiple data sources, including

mRNA expression, microRNA expression, number of copies variation and protein interaction data. For each subtype of ovarian cancer, the authors explored the oncogenic processes and the leading genes using a network-based approach. As results, they present a computational framework to harness the full potential of large-scale genomic data to discover the network modules of ovarian cancer subtypes and candidate genes for therapy.

Chang et al. [2011] use a Transcriptomic Signature Network for identification of lung cancer subtypes. During the study, they compared their model with a method using PCA-LDA (Principal Component Analysis and Linear Discriminant Analysis). As a comparison, they showed that their method achieves a maximum of 95.2 % accuracy while the PCA-LDA reaches 93.4%.

In multidimensional Fessler et al. [2016], the authors use a multidimensional network to identify groups of molecules responsible for each subtype of colorectal cancer. Thus, the authors defined a network-based approach that involves multiple molecular modalities such as gene expression, methylation levels, and microRNA expression (miR). The authors then showed that the determination of regulatory networks, groups of biomolecules that act as therapy, is a powerful strategy to define responsible groups of different subtypes of cancer because they have the ability to identify subtype affiliation and to define biological behavior.

In the work of Yang et al. [2017], the authors seek to better characterize the prostate cancer. For this, they used an approach of molecular networks and profiles of somatic mutations. The results of this characterization are compared with clinical and pathological results. The results obtained through molecular networks and mutation profiles indicate that prostate cancers can be classified according to their pro-patents of mutation and argue that these subtypes may help to improve the treatment of this type of cancer in the future.

To identify subtypes of breast cancer, Hua et al. [2013] construct a network of microRNA interactions where they apply a Silico method to perform the identification. The authors show that the microRNAs present excellent topological properties and are essential for unraveling their biological function. As results, they present a new Silico method to predict candidate microRNAs of breast cancer subtypes.

Dutta et al. [2012] construct networks to identify genetic networks responsible for subtypes of breast cancer. The authors are able to identify distinct genetic networks that were responsible for the three most common subtypes of breast cancer. Finally, they report that members of the triple-receptor-negative breast cancer (TNBC) genetic networks increased the functional specificity of TNBC cell lines and had a greater functional sensitivity compared to the genes selected only by differential expression,

facilitating the distinction between subtypes.

Mertins et al. [2016], which described proteogenic analyzes of breast cancer samples available in the TCGA database (The Cancer Genome Atlas) that represent the four major intrinsic subtypes of breast cancer. In one of their analyzes, the authors used cluster analysis and molecular networks, using Pearson correlation to construct their network. As a result, they demonstrate that the proteogenic analysis of breast cancer elucidates the functional consequences of somatic mutations, narrows the indications of the responsible genes in larger delimitations and identifies therapeutic targets.

2.4.2 Related Works Synthesis

In the work of Zhang et al. [2016], they built a network that represents a patient similarity network. The nodes are the patients, whereas the edges are weighted by similarity between patients. Chang et al. [2011] built a Transcriptional Network with the bayesian networks framework [Chang and Ramoni, 2009], in which a model encodes the dependence relation among the cancer subtype and genes. Yang et al. [2017] built a protein-protein network using STRING database [Szklarczyk et al., 2014], STRING Network, in which they use human protein-protein interaction data for network construction. miRNAs Interaction Networks are networks built from miRNA Expression Profiling, in which they infer large networks using mutual information, Hua et al. [2013] describe this approach to identify breast cancer subtypes. Finally, Dutta et al. [2012], Fessler et al. [2016], Mertins et al. [2016] built their networks applying Pearson Correlation and using their coefficients as edges. In this work, we call this approach of using correlation coefficients as edges as Weighted Correlation Network, described in Section 2.2.

The Table 2.1 show a Synthesis of the related works, based on their network approaches and the Cancer Type.

2.5 Chapter Discussion

This chapter presented a summary of fundamentals necessary for the work development. Our approach follows a Weighted Correlation Network, Network Analysis, and Biological concepts as tools to explore cancer characteristics.

Recently, with the advent of molecular biology, researchers can use gene expression data to reconstruct networks and find patterns that may be useful in discriminating some characteristics of cancer. They construct these networks and use the network

Title	Authors	Approach	Cancer Type
Identification of ovarian cancer subtype-specific network modules and candidate drivers through an integrative genomics approach	Zhang et al. (2016)	Similarity Networks	Ovarian
A transcriptional network signature characterizes lung cancer subtypes	Chang et al. (2011)	Transcriptional Network	Lung
A multidimensional network approach reveals microRNAs as determinants of the mesenchymal colorectal cancer subtype	Fessler et al. (2016)	WCN - Pearson	Colorectal
Molecular classification of prostate adenocarcinoma by the integrated somatic mutation profiles and molecular network	Yang et al. (2017)	STRING Networks	Prostate
Identifying breast cancer subtype related mirnas from two constructed mirnas interaction networks in silico method	Hua et al. (2013)	miRNAs Interaction Networks	Breast
A network-based, integrative study to identify core biological pathways that drive breast cancer clinical subtypes	Dutta et al. (2012)	WCN - Pearson	Breast
Proteogenomics connects somatic mutations to signalling in breast cancer	Mertins et al. (2016)	WCN - Pearson	Breast

Table 2.1: Related works synthesis.

topology to identify discriminative characteristics that can be used as characteristics vector of a biological method to be studied.

Therefore, we want to explore more network analysis than others works. We aim to find some characteristics useful for biology analysis, extract information about the network topology as we do in works with social networks for computational problems, imagining each gene as a node that also have value grouping with other genes.

In the next chapters, we present our approach, the results of our approach and conclusions.

3

Proposed Approach

In this chapter, we describe the proposed approach, explain the layers' correlation, the network structure, and explore network centralities and community detection to apply gene set enrichment analysis.

3.1 Overview

In Figure 3.1, we present an overview of our proposed solution. We divided in four steps: Correlation Tables, Weighted Correlation Network, Network Analysis, and Gene Set Enrichment Analysis.

In section 3.2, we show the first step of our approach, which consists in computing the correlation between lists of our Gene Expression Data. This step is important for section 2.2 in which we use the correlation coefficients to reconstruct our biological network structure, using the correlation as weight for our edges.

For our analysis, presented in two steps (section 3.4 and 3.5), we apply centrality measures and group genes in communities to analyze the topological information of the network. We gather information inherited by the centralities and the communities and use lists of genes as inputs to analyze the biological information using Gene Set Enrichment Analysis.

3.2 Correlation Tables

The data that we use consists in gene expression data. In our work, we use three main layers (CNA, RNA, and Protein). In each layer, we have columns representing the samples and lines representing genes (CNA and RNA layers) or proteins (Protein

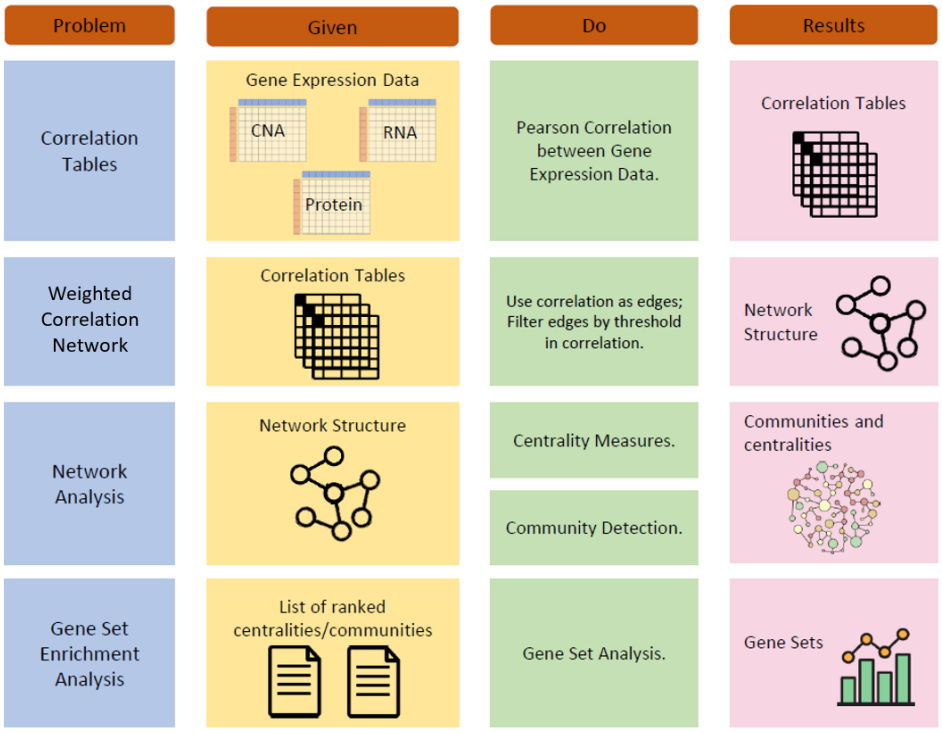


Figure 3.1: Proposed Solution.

layer), as shown in Figure 3.2. For our method, we compute the correlation between gene expression data, in order to explore the relations between genes and protein.

CNA

	sample 1	...	sample 77
gene 1			
...	...	...	...
gene n			

RNA

	sample 1	...	sample 77
gene 1			
...	...	...	...
gene m			

Protein

	sample 1	...	sample 77
protein 1			
...	...	...	...
protein k			

Figure 3.2: Layer's data.

3.2.1 Multi Layer Correlation

Multiomics refers to a biological analysis approach in which the data sets are multiple "omes", such as the genome (CNA), transcriptome (RNA), and proteome (Protein), in other words, the use of multiple omics technologies to apply in a study. In Multi

Layer Correlation, we correlate all table data in the multiomic order (CNA ↔ RNA ↔ Protein), as shown in Figure 3.3. The definition of multiomic The purpose of this correlation is to show the behavior of our multi layer network, connecting all three layers of gene expression data. After the correlation process, we build two tables of data correlation coefficients, CNA-RNA correlations and RNA-Protein correlations to construct one network.

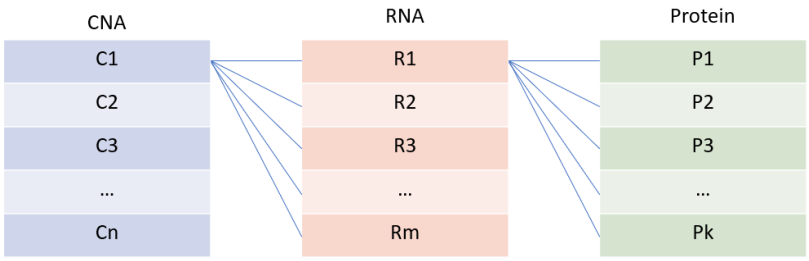


Figure 3.3: Correlate gene expression data between CNA, RNA and Protein.

3.2.2 Single Layer Correlation

In Single Layer Correlation, we correlate only the layers that can possibly induce his kind, RNA and Protein. This correlation is important to show the interrelationship between genes and proteins of same layer (RNA ↔ RNA and Protein ↔ Protein), shown in Figure 3.4. CNA does not have a network because Copy Number Alteration does not have relationship with its kind. At the end, we build two tables of data correlation coefficients, RNA-RNA correlations and Protein-Protein correlations to construct two networks.

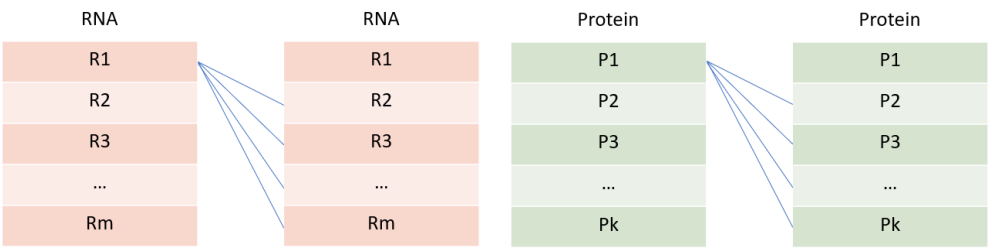


Figure 3.4: Correlate gene expression data between RNA with RNA, and Protein with Protein.

3.3 Weighted Correlation Network

After we have the tables of data correlations coefficients, we build a network where the genes/proteins are nodes and the correlation values are the weighted edges. To look for more significant networks we filter the correlations by choosing threshold values. In Figure 3.5, we could see a sketch example of connected nodes and disconnected nodes from CNA, RNA, and Protein. In the example, the disconnected node 'p1' was filtered by a threshold. Our edges values are correlation modules because we look for correlation regardless the type: positive or negative. Besides that, this simplification works better for computing centralities. Most of the centrality measures work with positive edges, when we have a loop of negative edges the measures may fall in infinite loop, having problems with their computations.

Once we defined our edges values, we apply the threshold to build the network. This value can change by looking the different patterns shown when we apply network centrality measures and community detection. Reconstructing a network is the first step of our configurations, the correlation coefficients may not change but the network structure can change by tuning the threshold.

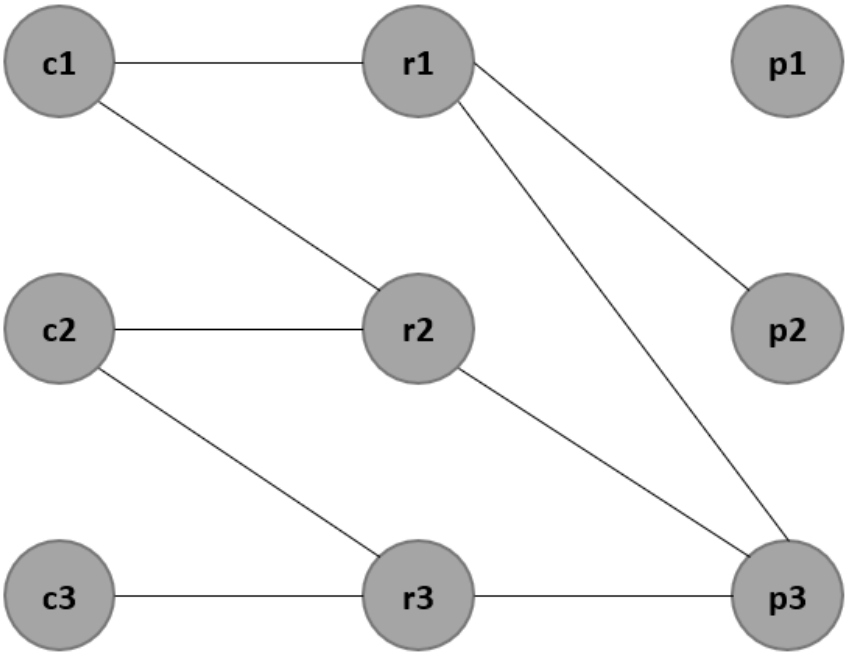


Figure 3.5: Network structure.

3.4 Network Analysis

Given the WCN built, we apply network centrality measures and community detection. We apply centrality measures to evaluate influential nodes. The different edges' values will give different influential nodes, so we generate different ranks depending on the centrality. In community detection, we can see how a group of nodes interact, which clusters are formed, which node is the most influential one inside each community, etc.

Community Detection and Centrality Measures may not be a novel approach in cancer analysis, but there are plenty of combinations using them that were not tested yet. Our main purpose is to clarify important nodes, grouped or not, that have meaning in the cancer characteristics. For that goal, we explore these network analysis through comparison between different cancer data, removal of important groups to see how important they were for the network, and others combinations that may show us some meaningful importance of groups and individual nodes.

3.5 Enrichment Analysis

Once we have lists of genes ranked by centrality measures or grouped by communities, we apply enrichment analysis to interpret our data. Thus, we can identify important gene sets related to a group or a ranked list generated by our strategy.

To identify a gene set, in a collection of annotated gene sets, related to our list is an important task. We can define the gene sets found as a characteristic for our specific type of cancer.

3.6 Chapter Discussion

Each step of our approach have some configurations to decide: (a) the threshold that filters the correlations for our Weighted Correlation Network; (b) centrality measures used in our study; (c) community detection method applied and (d) collection of gene sets compared.

After we decided each configuration for our studies, we approached our network-based exploration using enrichment analysis. The experimental methodology and the results of our exploration are described in chapter 4.

4

Experiments

In this chapter we show the experimental methodology, some experiments and results achieved.

4.1 Experimental Methodology

Our database was provided by researcher collaborators from NYU Medical School. It contains information about proteogenomic data from Breast cancer and Ovarian cancer. The Breast cancer and Ovarian cancer data have 4 layers: CNA, RNA, and Protein.

4.1.1 Gene Expression Data Format

A gene and protein expression can be represented by a real-valued. In this work, the values presented in each layer are normalized using Z-score transformation with average 0 [Cheadle et al., 2003]. Our dataset is divided in tables, where each table organizes data into m columns (samples) and n rows (genes, proteins).

4.1.2 Preprocess

In our project, we work with CNA, RNA, and Protein, this is because of the highly missing data occurred in the Phosphoprotein layer. The miss information is a problem for our project, so we decided to use only variables with 30% or less missing data for each pairwise in correlation method. After we decided which layers would be used and how the correlation method would deal with missing data, we cleaned the data. The cleaning was based on samples that appear in the three layers (CNA, RNA, and Protein) because we needed each sample as participant in all layers of our correlation

tables. For example, samples that have only RNA gene expression data and did not appear in both other tables. After this cleaning process, we ended with 77 samples for Breast cancer data, and 173 samples in the Ovarian cancer data.

4.1.3 Correlation Tables

To construct correlation tables necessary for our networks, we applied Pearson correlation (Section 2.2) for both Breast Ovarian cancer. The choice of the correlation was based on the most used correlation in the related works presented in Section 2.4.

4.1.4 Weighted Correlation Network and Network Analysis

In all our experiments, we used R programming, to calculate all correlation coefficients with respective p-values, to reconstruct our networks, and to apply network analysis. For our network visualization, we used the software GEPHI [Bastian et al., 2009].

4.1.5 Enrichment Analysis

Identifying a gene set related to our lists of rankings or communities is an important task to define characteristics for cancer. Different enrichment analysis methods can be used for different inputs, depending on the information given in the input lists. In this work, we used two enrichment analysis methods: PANTHER Classification and GSEA PreRanked. Those enrichment analysis methods can work with list of genes sorted, in GSEA PreRanked, or not, in PANTHER Classification, to identify biological related set of genes.

PANTHER Classification

PANTHER Classification System [Mi et al., 2019] evaluates the lists of genes that belong to a given gene or protein family or subfamily, have a given molecular function or participate in a given biological process or pathway.

PANTHER Classification uses a statistical test method (Fisher's exact test) [Raymond and Rousset, 1995] for the PANTHER overrepresentation test. It consists in compute whether the proportions of a gene list given is significantly present in gene sets on PANTHER database.

We use this enrichment analysis in community detection gene lists because we do not have ranked list of genes, this system only needs a list to match genes in specific biological backgrounds.

GSEA PreRanked

Among the most used methods for Enrichment Analysis, GSEA (Gene Set Enrichment Analysis) [Subramanian et al., 2005]. It computes the Enrichment Score (ES) that is a score given to a gene set when a ranked list is matching its genes. GSEA PreRanked method is able to calculate all enrichment scores (ES) for all gene sets. These ES are scores that increase when, in a ranked list of genes, each sorted element is found in the ranked list or studied gene set in database, showing overrepresented extremes of the gene list when it occurs. The extremes of the gene list are called "*na_pos*", when in the top of list, and "*na_neg*", when in the bottom of list.

To measure each ranked list, the method computes an enrichment score (ES) that reflects the degree to which a gene set S is overrepresented at the extremes (top or bottom) of the entire ranked list L. The score is calculated by walking down the ranked list L, increasing a running-sum statistic when the method encounters a gene in S and decreasing it when encounters genes not in S [Subramanian et al., 2005]. In this work, we focus on quantifying the gene sets found as enriched through ES. Hence, GSEA method provides us a False Discovery Rate (FDR) value for each gene set.

The false discovery rate (FDR) is the estimated probability that a gene set with a given enrichment represents a false positive finding. For example, an FDR of 25% indicates that the gene sets found are likely to be valid 3 out of 4 times. The GSEA analysis reports highlights enrichment gene sets with an FDR of less than 25% as those most likely to generate interesting hypotheses of related gene sets and drive further research, but provides analysis results for all analyzed gene sets.

The FDR is a ratio of two distributions: (1) the actual enrichment (ranked list enrichment) versus the enrichment scores for all gene sets against all permutations of the dataset and (2) the actual enrichment (ranked list enrichment) versus the enrichment scores of all gene sets against the actual dataset. For example, if you analyze four gene sets and run 1000 permutations, the first distribution contains 4000 data points and the second contains 4. In our work, we use the FDR as gene set reliability control for each gene set, that is the probability that our ranked list is wrong enriched for that set.

This method was used in centrality measures gene lists because we have a ranked list of genes. GSEA PreRanked only needs an ordered list to match genes in specific biological backgrounds, matching them in a collection of annotated gene sets.

4.2 Multi Layer - CNA/RNA/Protein

In this experiment, we connect multiomic gene expression data. This is, connect CNA to RNA and RNA to Protein, following the interactions that represents the central dogma of molecular biology [Crick, 1970]. This experiment is important to reconstruct a molecular network that shows CNA, RNA, and protein layers.

The purpose of this experiment is to show the behavior of our multi layer network when we apply community detection. From our knowledge in network science, this method is a process of discovering cohesive groups in a network. The objective applying this method is to reveal if the network is grouping genes and proteins from different layers as a biological cohesive group.

In the following subsections, we present the Weighted Correlation Network, choosing better thresholds to build our network, the network analysis, applying community detection and comparing communities, and then a biological analysis, which consists in evaluating highlighted communities.

4.2.1 Weighted Correlation Network

In Figure 4.1 and Figure 4.2, we observe the distribution of Breast Cancer correlation coefficients for CNA-RNA and RNA-Protein. In Figure 4.3 and Figure 4.4, we observe the distribution of Ovarian Cancer correlation coefficients for CNA-RNA and RNA-Protein. For networks analysis, we must keep a trade-off between dense network with low correlations (lower thresholds) and sparse network with high correlations (higher thresholds). To this end, we choose moderate to high correlation coefficients (0.5, 0.6 and 0.7) [Mukaka, 2012].

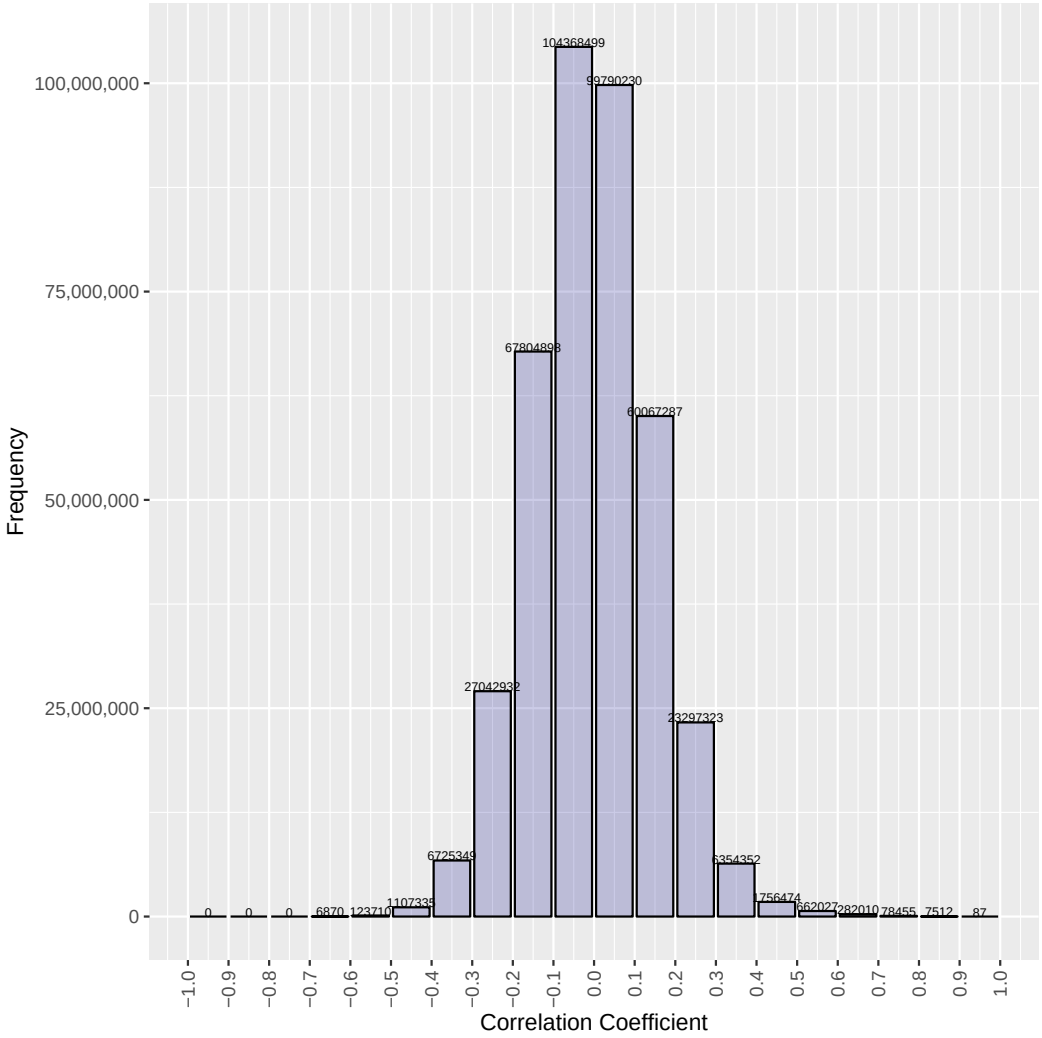


Figure 4.1: Correlation distribution for Breast CNA-RNA Gene Set Data.

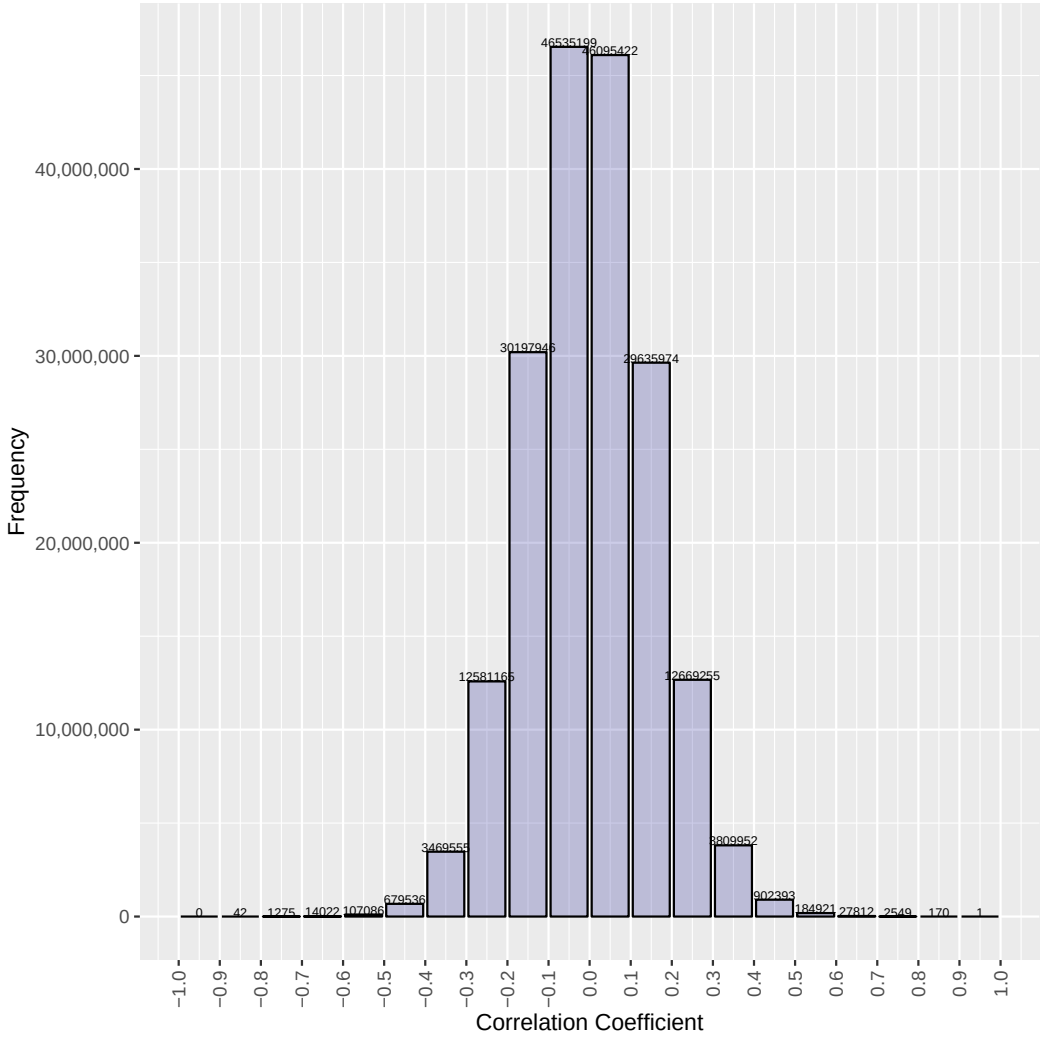


Figure 4.2: Correlation distribution for Breast RNA-Protein Gene Set Data.

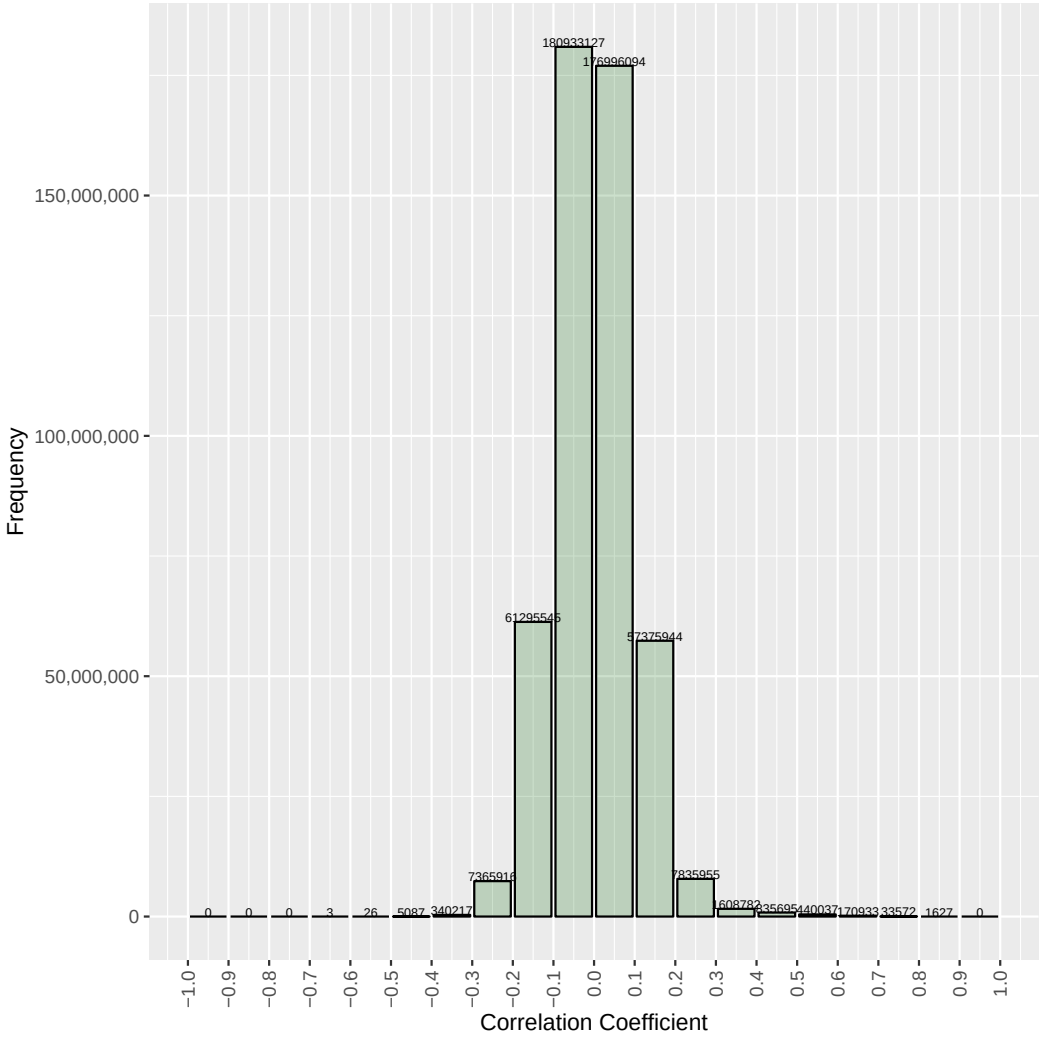


Figure 4.3: Correlation distribution for Breast CNA-RNA Gene Set Data.

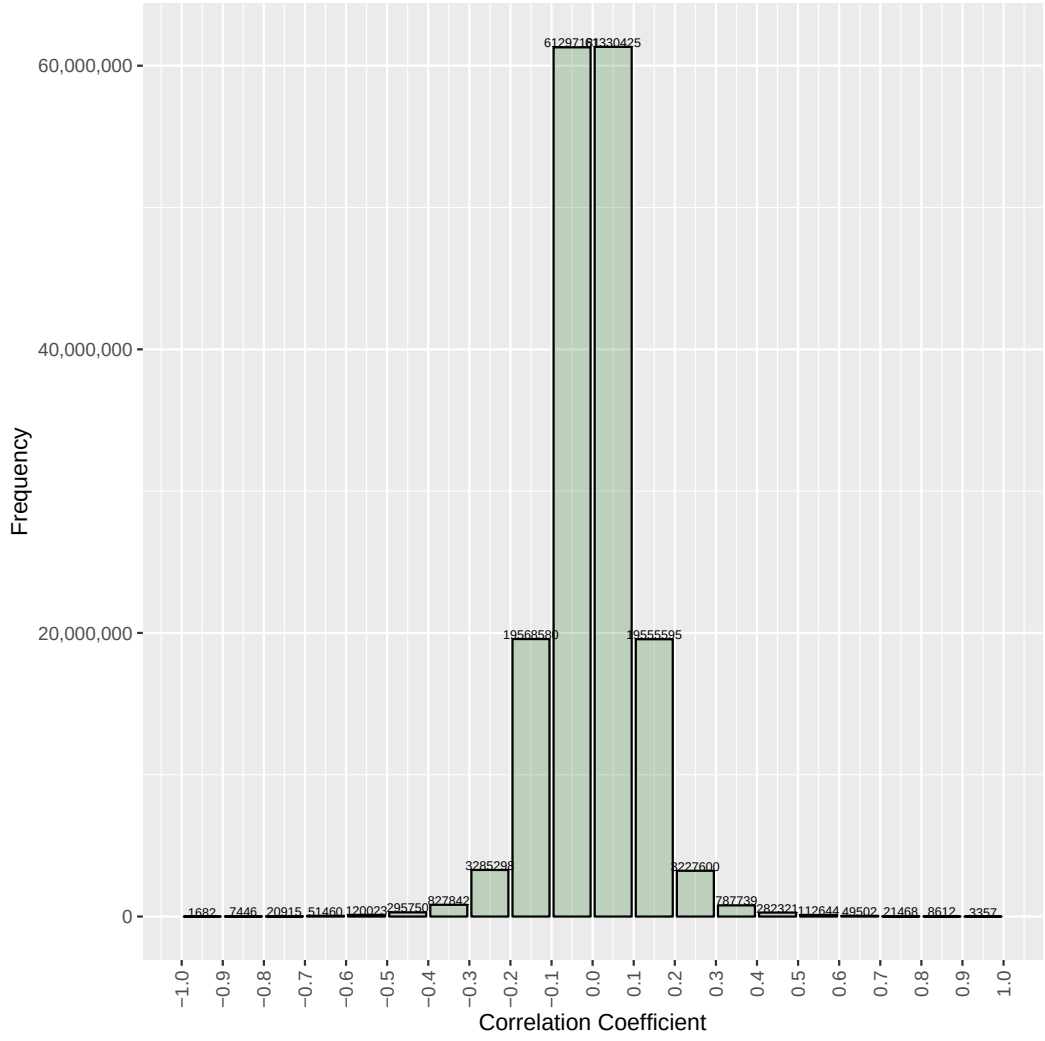
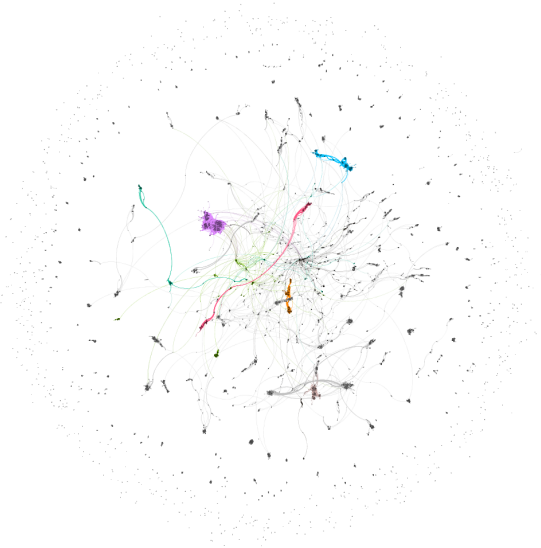


Figure 4.4: Correlation distribution for Ovarian RNA-Protein Gene Set Data.

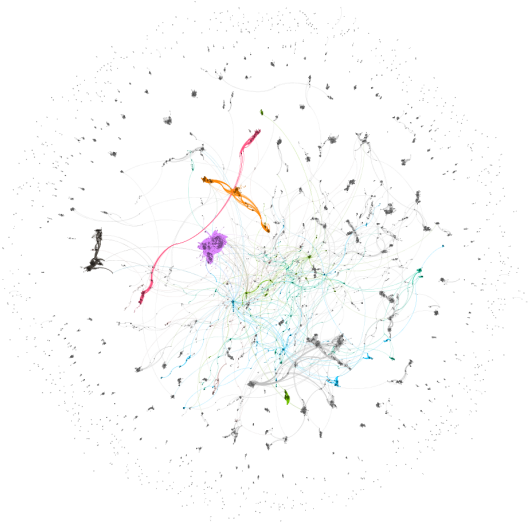
4.2.2 Network Analysis

The purpose of this experiment is to study if we can capture biological processes from a network built using all layers when we apply community detection.

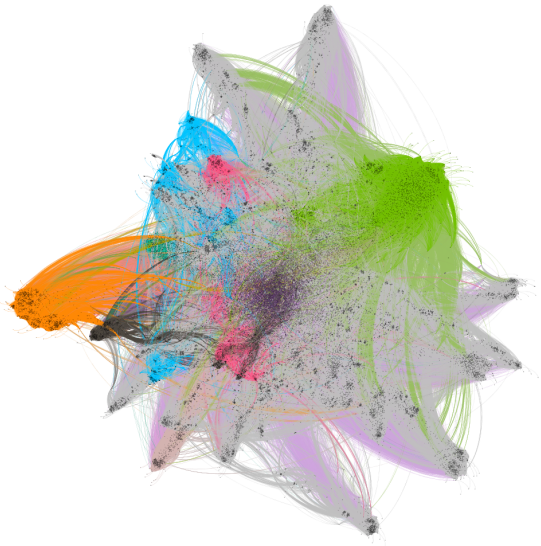
For Breast Cancer, we can see that the first two networks (Figures 4.5a and 4.5b) do not show dense communities as the third network (Figure 4.5c), this could be result of the high number of new connections appearing from threshold 0.6 to 0.5.



(a) Network for threshold 0.7.



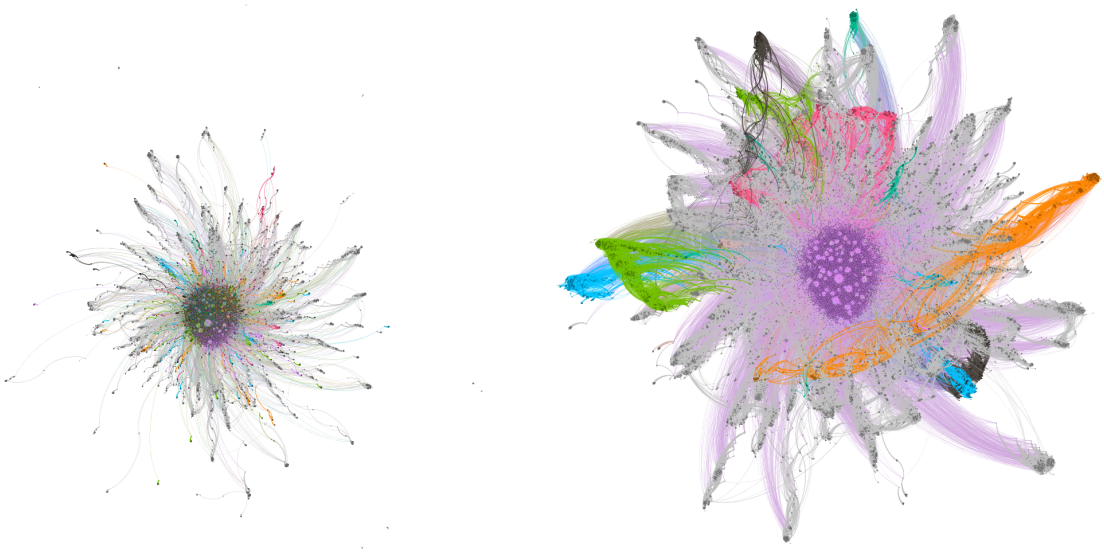
(b) Network for threshold 0.6.



(c) Network for with threshold 0.5.

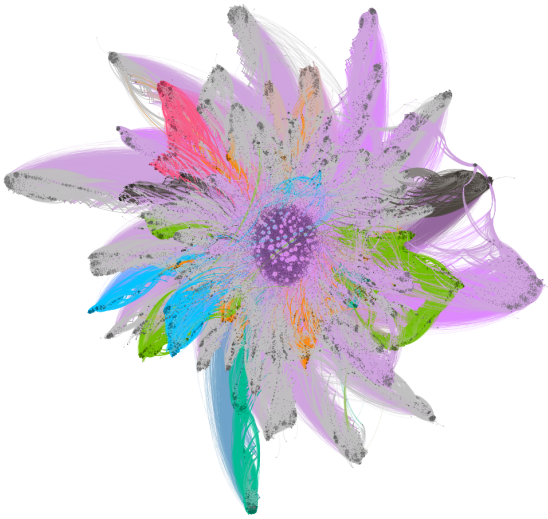
Figure 4.5: Breast Cancer Networks.

For Ovarian Cancer, we can see a different outcome right in the first network (Figure 4.6a). Although we do not have high dense communities, we see that a central community starts to gather nodes. From the network 0.6 (Figure 4.6b) to network 0.5 (Figure 4.6c), visually we do not see a great difference from the threshold decrease, but we see that the central community remains. Further explorations shows that the central community is the biggest one.



(a) Network for threshold 0.7.

(b) Network for threshold 0.6.



(c) Network for with threshold 0.5.

Figure 4.6: Ovarian Cancer Networks.

To monitor the communities as we change threshold, Figure 4.7 shows how communities behave through thresholds on both network cancers. We can highlight that as we lower the threshold, the number of communities decreases as the larger community emerges. The reason is that the number of edges increase, connecting more nodes and creating new paths between nodes.

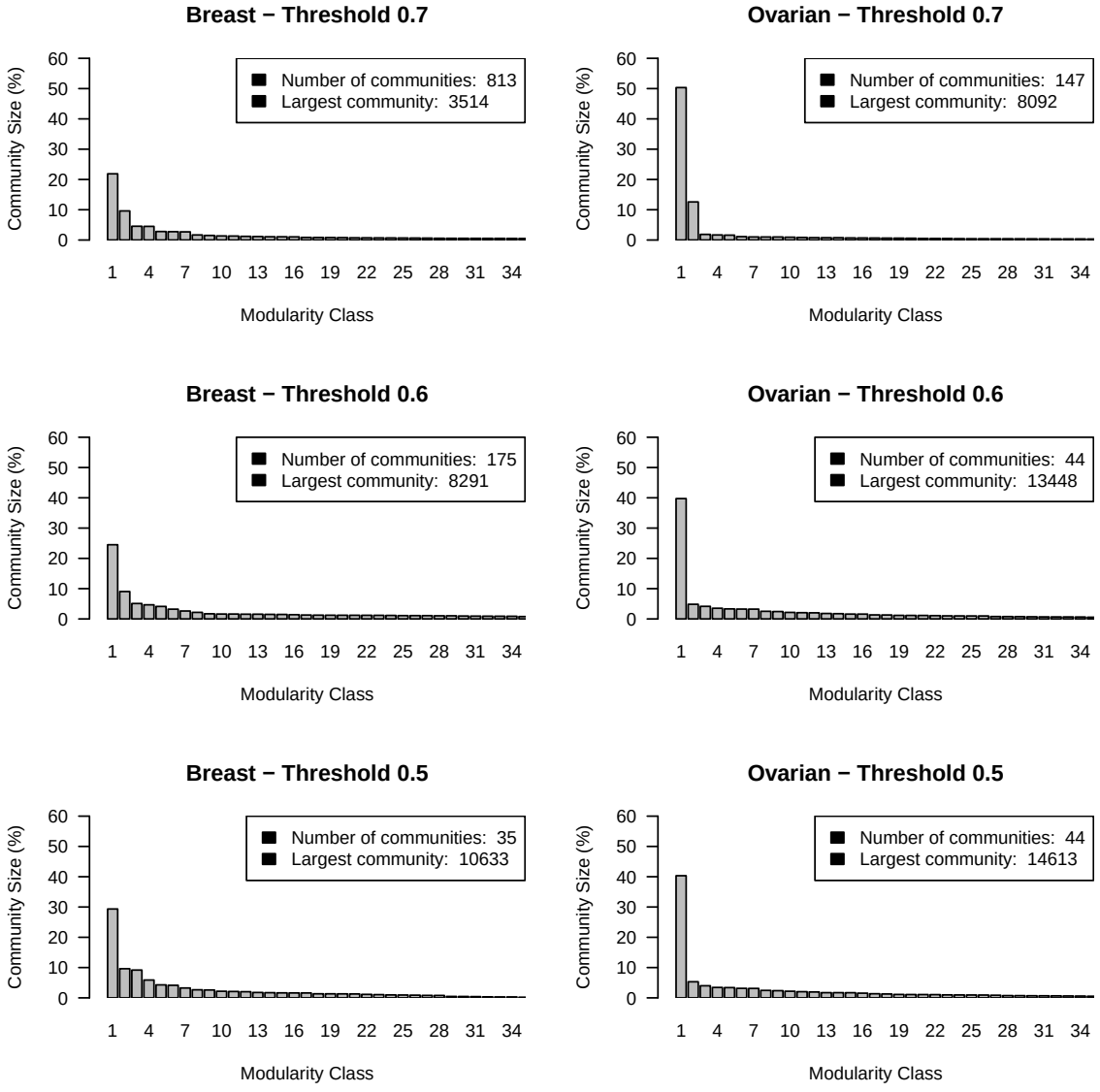


Figure 4.7: Distribution of Community Size.

As a way of community analysis, we compare community intersections between Ovarian and Breast cancer networks. For the study of the community intersections, we used a R library called Circlize [Gu et al., 2014]. This tool can generate a circular chart, where we can connect different sides of the chart and show the intersections occurring between both sides. We exclude genes and proteins that appear in only one

network (Breast or Ovarian) and identify genes and proteins that are in communities from both networks.

In this comparison, each side of the chart contains the communities of a network of a different cancer (Breast and Ovarian). We compared same thresholds of different cancer, lowering the threshold in each comparison. When we lower the threshold below 0.6, we can see a highly dense community intersection between Breast and Ovarian cancer.

Comparing Figures 4.8 and 4.9, we are able to see the difference as we lower to Figure 4.10, comparing the intersections presented with threshold 0.5 and 0.6, the biggest community intersection looks more consistent than networks with threshold 0.7.

In Figure 4.9 we clearly see a difference with their community intersection. A big intersection between biggest communities in Breast and Ovarian began to emerge.

In Figure 4.10 we can see the same behavior shown on Figure 4.9. The biggest intersection started to stand, gathering more intersected genes and proteins from both networks.

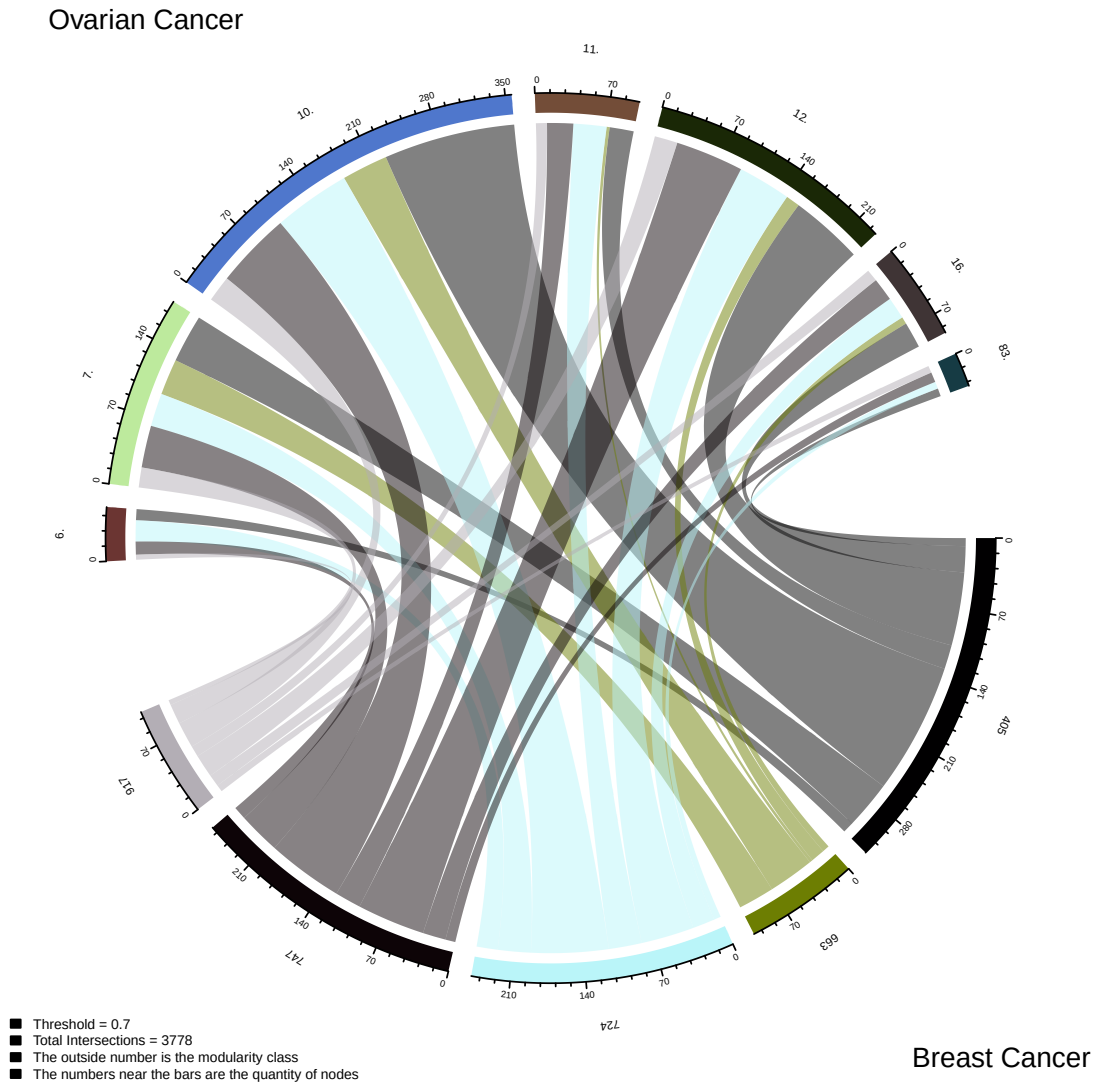


Figure 4.8: Community intersection between Ovarian and Breast cancer with threshold 0.7.

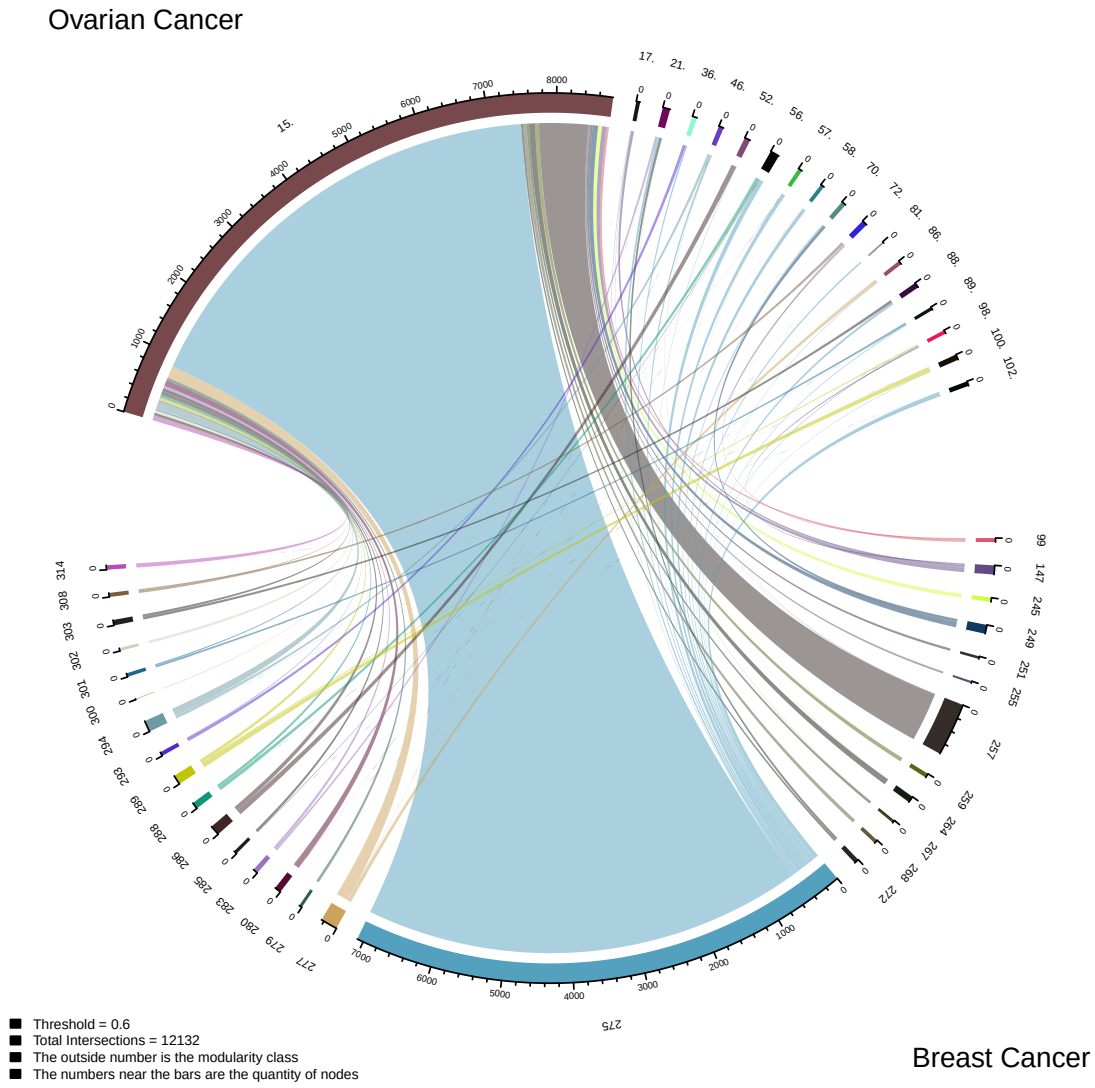


Figure 4.9: Community intersection between Ovarian and Breast cancer with threshold 0.6.

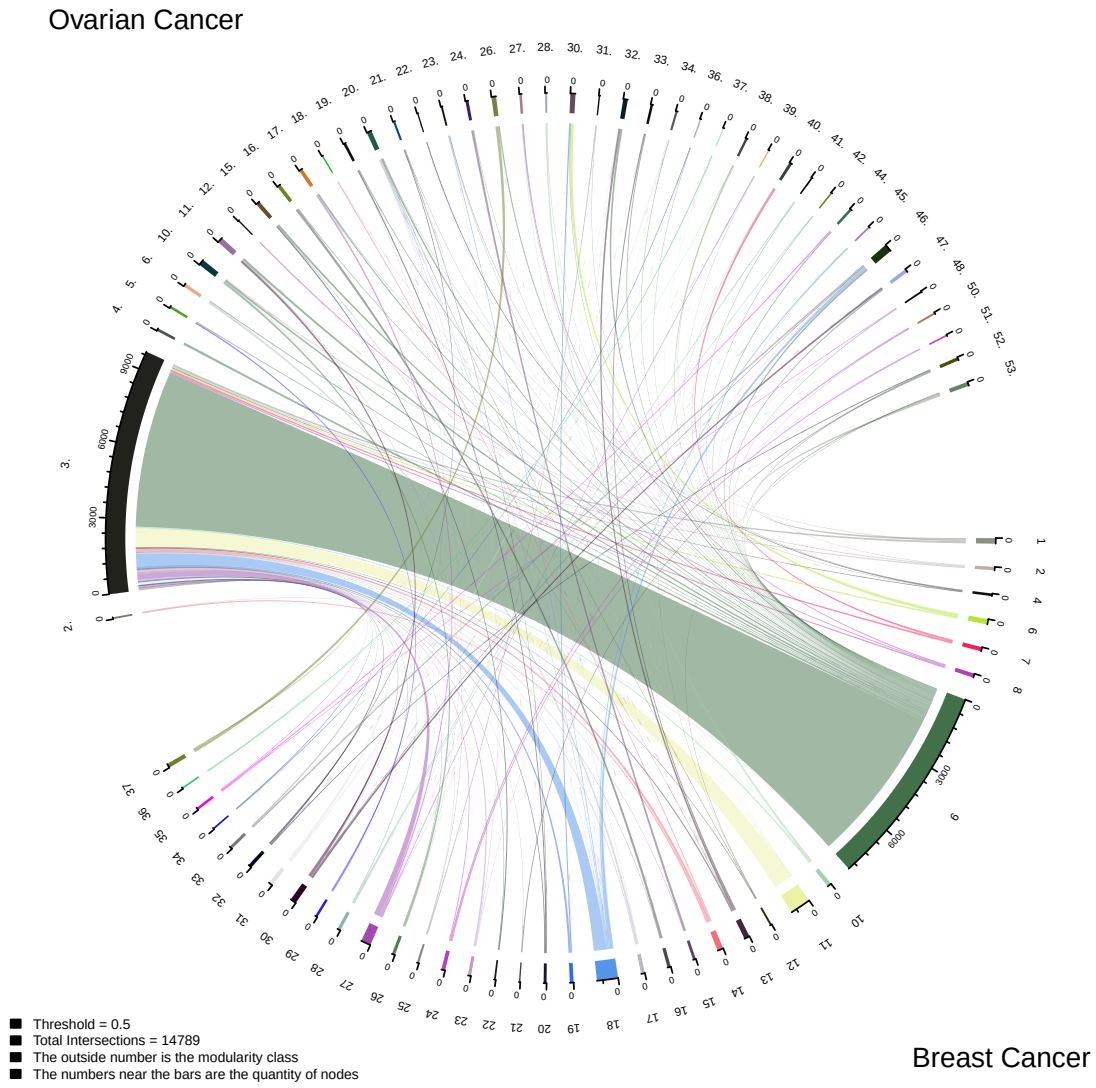


Figure 4.10: Community intersection between Ovarian and Breast cancer with threshold 0.5.

4.2.3 Enrichment Analysis

For this analysis, we took the biggest intersection between the communities from Breast and Ovarian networks shown in Figure 4.10. From this intersection, we separate two lists, one for RNA and other for Protein. CNA did not show expressive genes in the intersection between the biggest community.

For the RNA analysis, shown in Figure 4.11, we took a total of 6,162 genes inside the intersection from networks with threshold 0.5. From our analysis, we can highlight two biological processes expressed in dark blue and gray (two talest bars). Those two biological processes are responsible for cellular process and metabolic respectively.

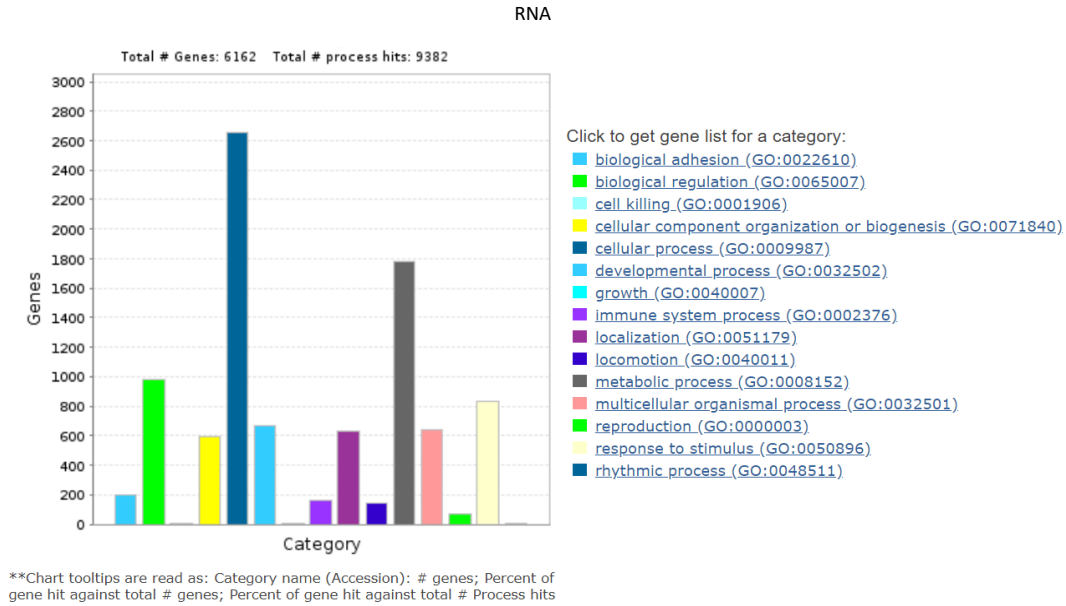


Figure 4.11: PANTHER analysis for RNA list.

For the Protein analysis, shown in Figure 4.12, we took a total of 365 proteins inside the intersection from networks with threshold 0.5. We can also highlight the same two biological process expressed in dark blue and gray (two longest bars). Those two biological processes are responsible for cellular and metabolic processes, respectively. Those similar results between RNA and Protein, having same two biological processes intersections, shows that our networks are mirroring the functioning of a cell. The reason is that those processes need either RNA and Protein for their functioning, and our community capture the biological analysis for both.

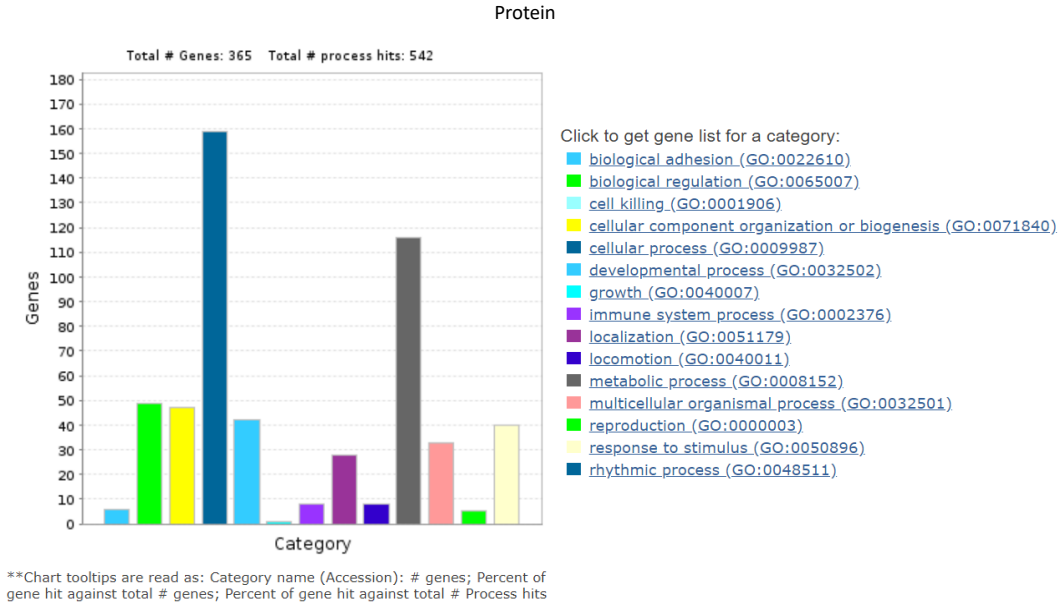


Figure 4.12: PANTHER analysis for Protein list.

4.3 Single Layer - RNA/RNA and Protein/Protein

In these experiments, we connect single layer gene expression data. This is, we connect data from same layer, connecting RNA to RNA in one experiment and Protein to Protein in the other. This experiment is important to show the interrelationship between genes and proteins of same layer, experiments in other works also relate biomolecules of same layer. CNA does not have a network because CNA genes do not have a biological relation, connecting only with RNA as we did for the multi layer experiment.

This experiment analyzes a set of ranked lists of RNA-RNA and Protein-Protein applying centrality measures as a way of sorting genes and proteins. Those ranked lists will serve as inputs to enrichment analysis methods which need ranked genes for gene set enrichment.

4.3.1 Weighted Correlation Network

The Figures 4.13 and 4.14 are the distribution of Breast Cancer correlation coefficients for RNA-RNA and Protein-Protein. The Figures 4.15 and 4.16 are the distribution of Ovarian Cancer correlation coefficients for RNA-RNA and Protein-Protein. We learned from last experiment that we must keep a trade-off between dense network with low correlations (lower thresholds) and sparse network with high correlations (higher thresholds). For this reason, we choose moderate correlation coefficients (0.5, 0.6 and 0.7) [Mukaka, 2012].

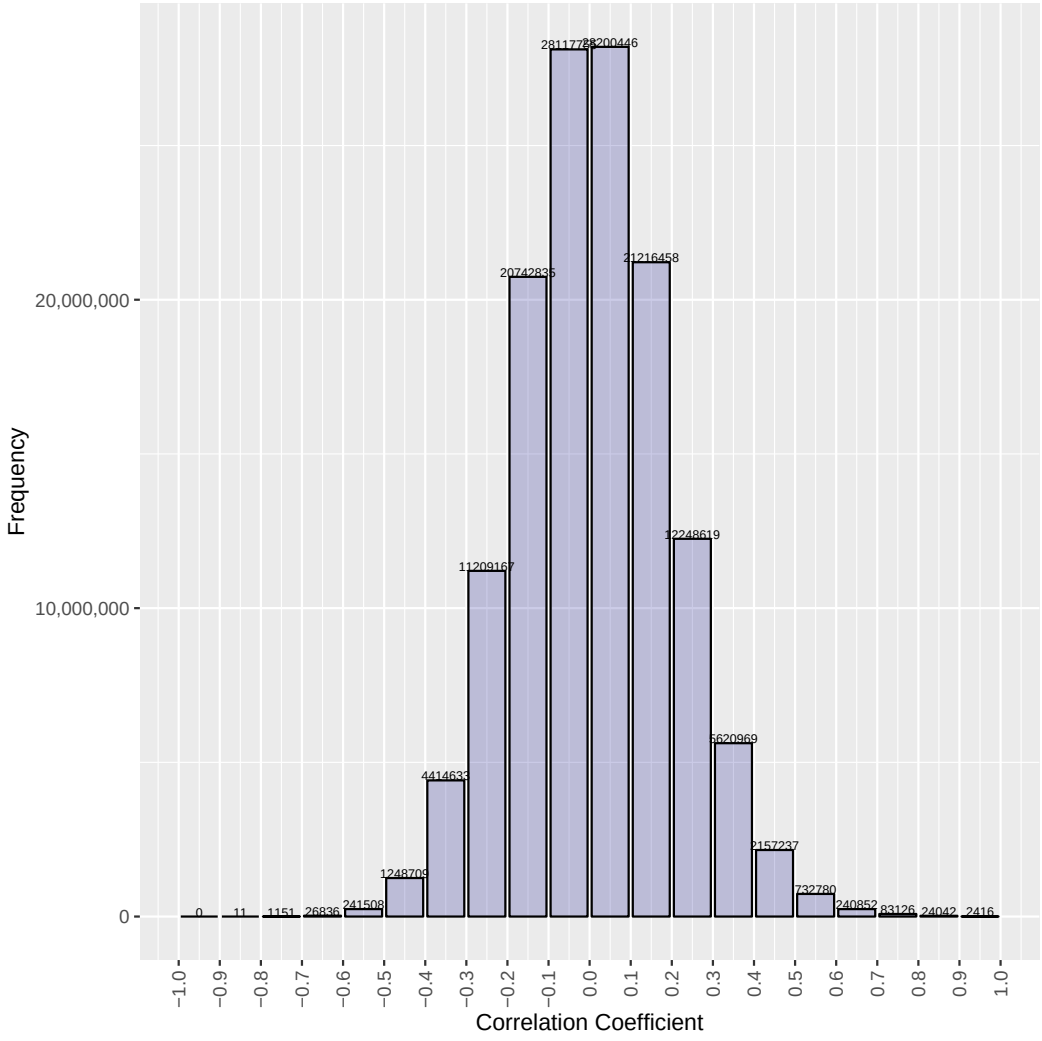


Figure 4.13: Correlation distribution for Breast RNA-RNA Gene Set Data.

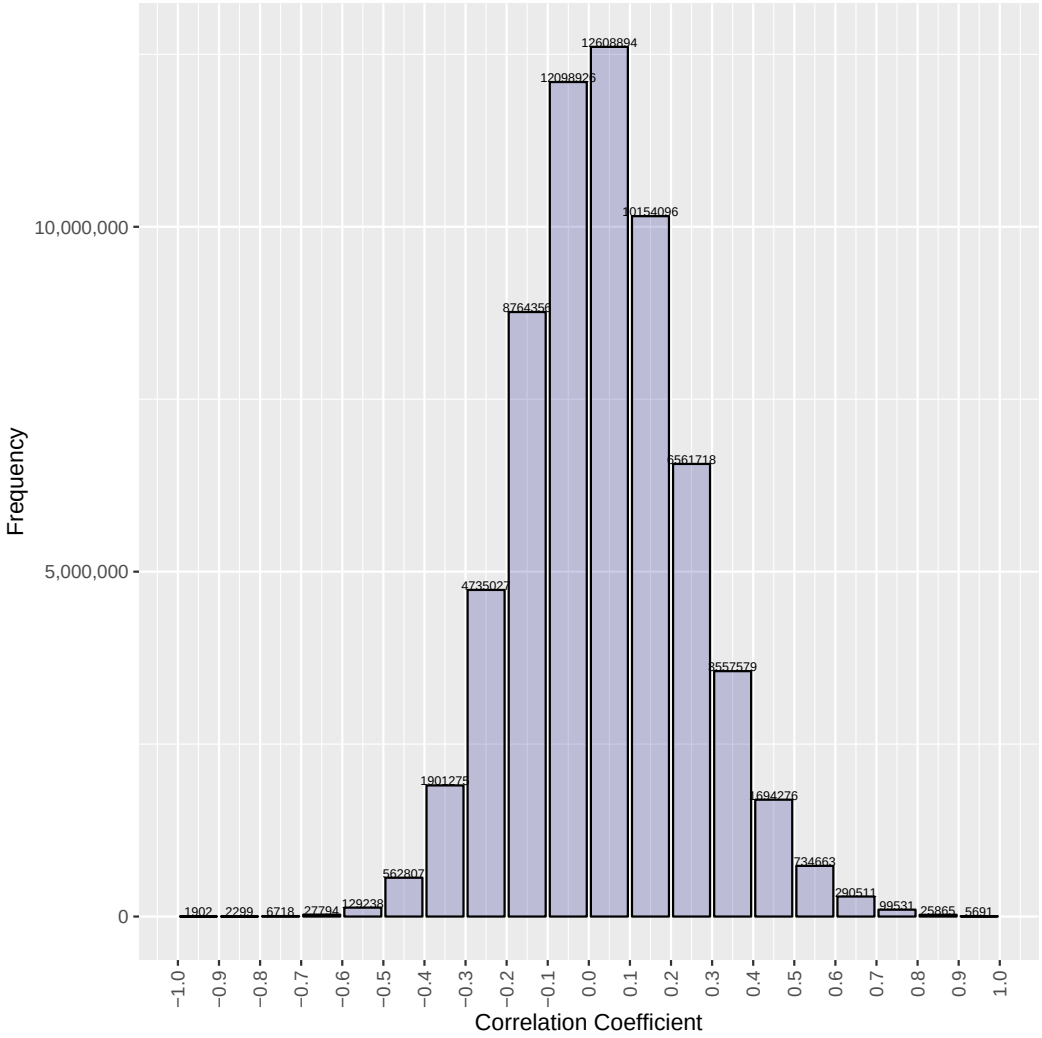


Figure 4.14: Correlation distribution for Breast Protein-Protein Gene Set Data.

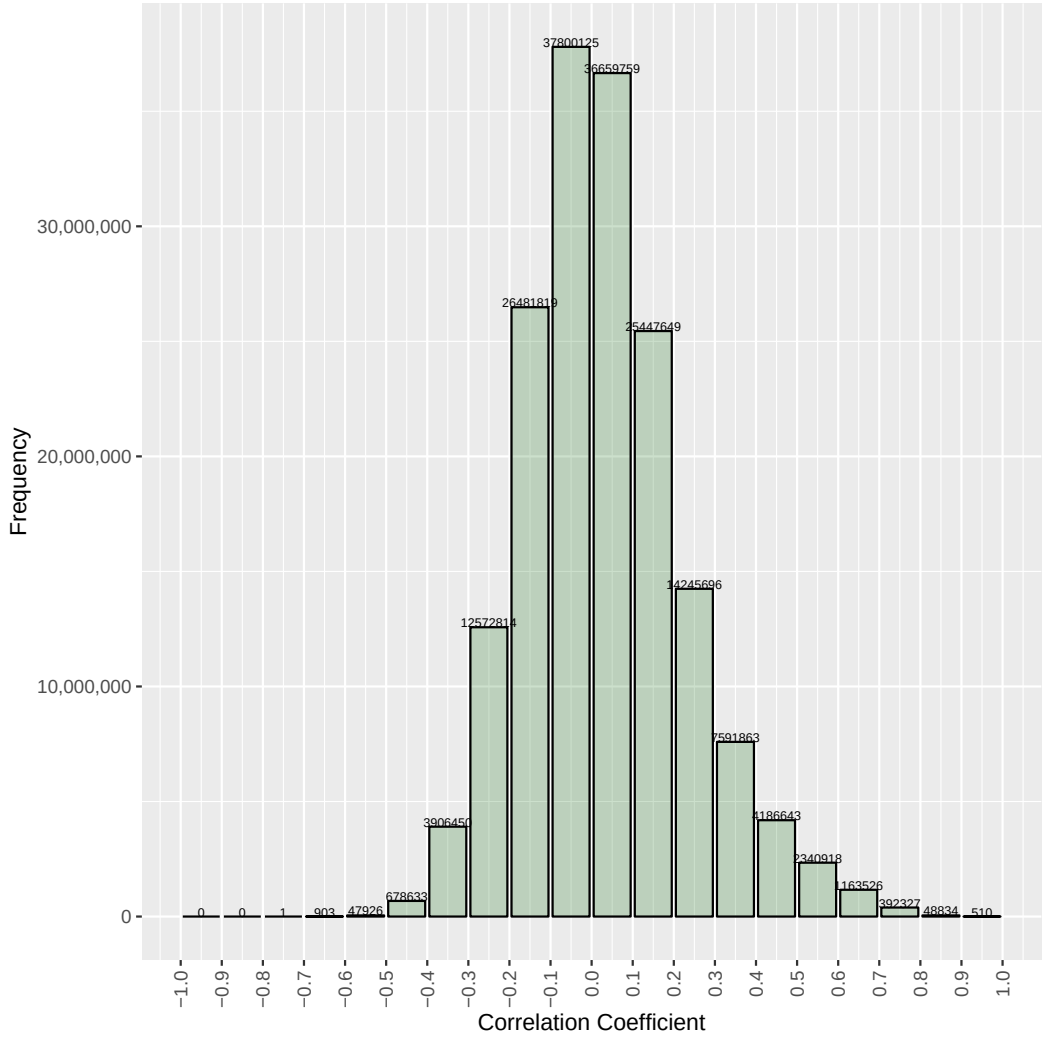


Figure 4.15: Correlation distribution for Ovarian RNA-RNA Gene Set Data.

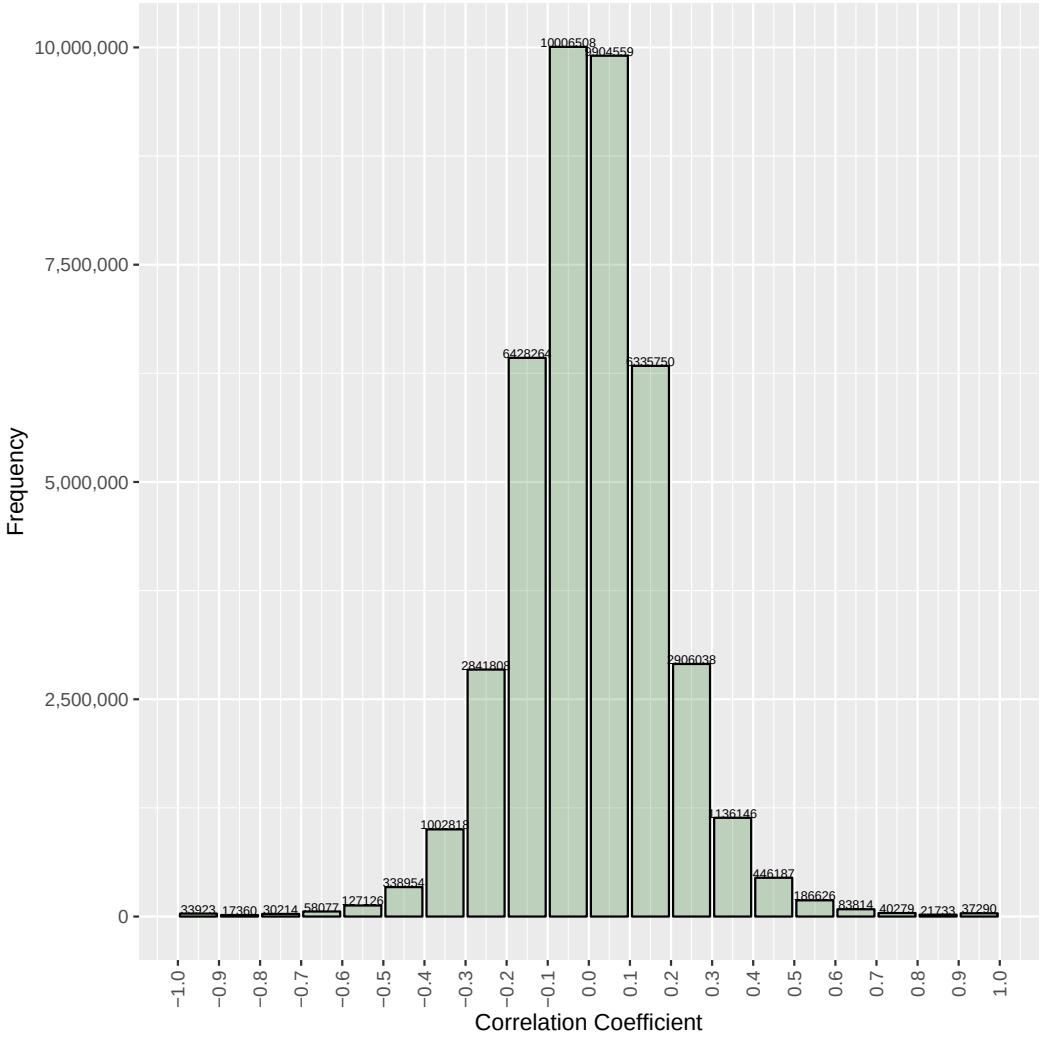


Figure 4.16: Correlation distribution for Ovarian Protein-Protein Gene Set Data.

4.3.2 Network Analysis

Our purpose here is to analyze the values of centralities, using their measures as a ranked list for enrichment analysis. In this step, we applied all centrality measures shown in Chapter 2. It is important to follow the number of experiments shown in next step, because we cross correlations with different networks.

We have a total of twelve networks as shown in Figure 4.17. There are 4 types of network: Breast Protein, Breast RNA, Ovarian Protein and Ovarian RNA; and 3 correlation coefficients: 0.5, 0.6 and 0.7. We apply all 10 centralities for those twelve networks, which provide us a total of 120 ranked lists for enrichment analysis.

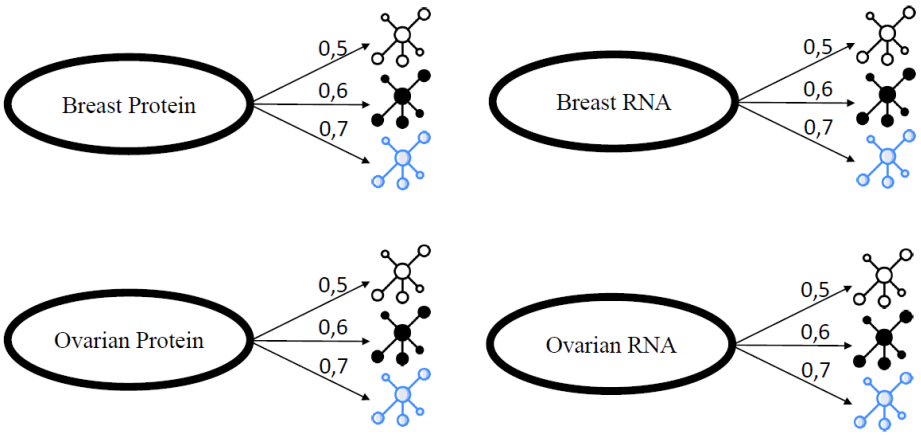


Figure 4.17: Networks submitted to analysis, we have 4 types of networks and 3 correlations.

For experiments in Enrichment Analysis, we set a maximum size of the ranked list of 5,000 elements and removed all nodes with no connections. In Table 4.1, we show the amount of elements left for enrichment analysis. We can see that most of the networks reach the limit of elements but three networks, in RNA, are under 5,000 elements. These are: Breast RNA 0.6 and 0.7, and Ovarian RNA 0.7.

4.3.3 Enrichment Analysis

In this step, we analyze gene sets in our 120 ranked lists using GSEA method. The Molecular Signatures Database (MSigDB) has annotated gene sets for use with GSEA software. In our analysis, we use all gene sets from MSigDB, a total of 17,810 gene sets. Then, we label gene sets, as shown in Table 4.2, as *"not cancer"*, *"cancer"*, *"breast cancer"*, and *"ovarian cancer"* gene sets based on their description. For this, we implemented a website crawler to get a description about each gene set and ran a script to label as follows: (1) *"cancer"* for keywords *cancer*, *cancinoma* or *tumour*; (2) as *"breast*

Network Type	Correlation	Quantity of Elements
Breast Protein	0.5	5,000
	0.6	5,000
	0.7	5,000
Breast RNA	0.5	5,000
	0.6	3,754
	0.7	1,861
Ovarian Protein	0.5	5,000
	0.6	5,000
	0.7	5,000
Ovarian RNA	0.5	5,000
	0.6	5,000
	0.7	4,175

Table 4.1: Ranked List size for each Network Type and Correlation.

Gene Set Label	Keywords	Quantity of Gene Sets
Cancer	{cancer, carcinoma, tumour}	3,079
- Breast Cancer	{breast, mammary, mamma}	582
- Ovarian Cancer	{ovarian, ovaries, ovary}	73
Not Cancer	not in cancer	14,731

Table 4.2: Gene Set Labels showing Keywords and Quantity.

cancer" for *"cancer"* keywords and *breast*, *mammary* or *mamma*; (3) *"ovarian cancer"* when inside *"cancer"* and for keywords *ovarian*, *ovaries* or *ovary*; and (4) *"not cancer"* when out of all previous sets. Based on our labels, we divided MSigDB in 3,079 gene sets related to cancer, 582 related to Breast Cancer gene sets, and 73 related to Ovarian Cancer gene sets.

We present one histogram for each type of network, as shown in Figure 4.18, in which: (1) y axis presents the trusted gene sets, which are gene sets with FDR (False Discovery Rate) less than 0.1; (2) x axis presents the centrality measures applied; and (3) the color of each bar represents the correlation coefficient for that network, showing three sets of bars for each centrality.

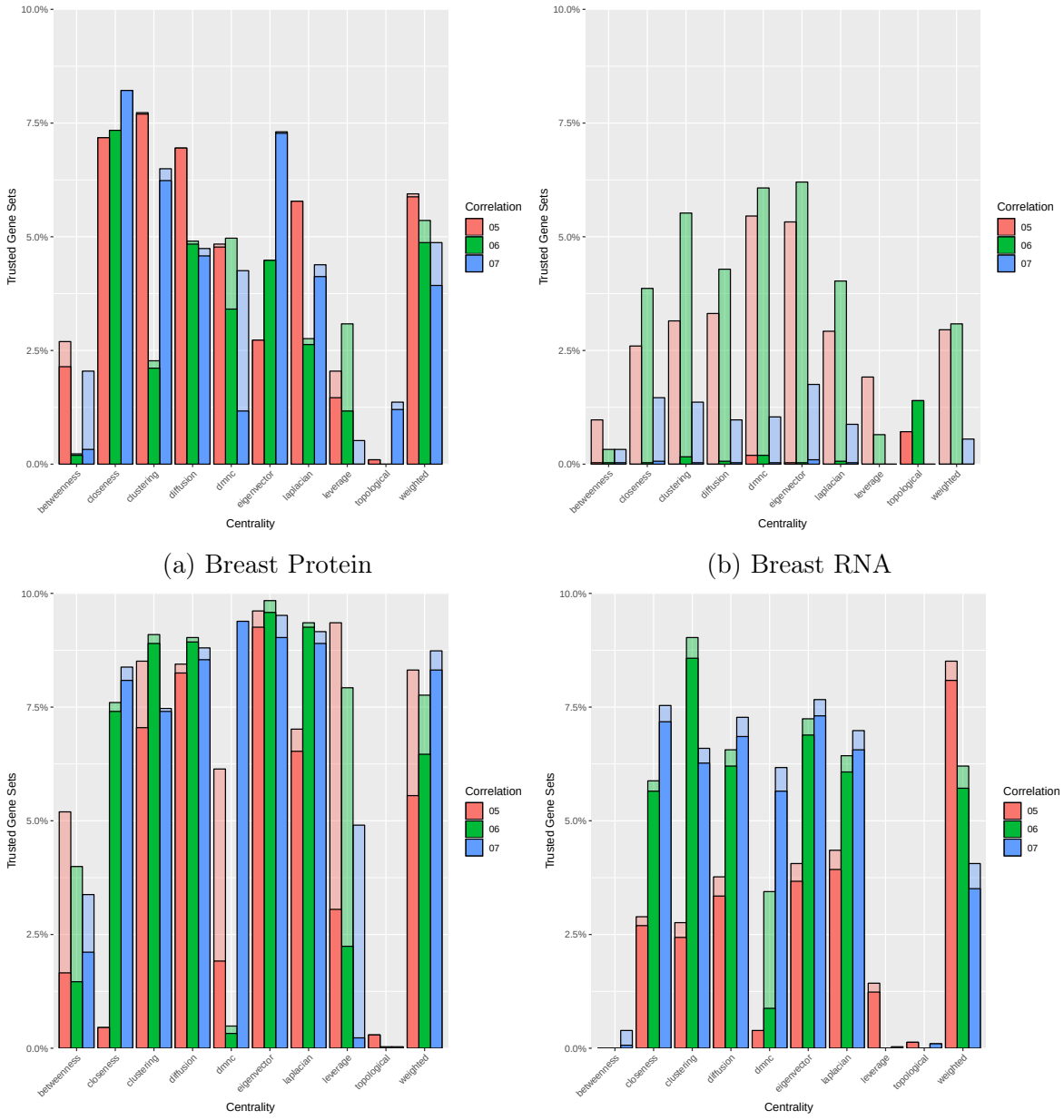
Cancer Only Gene Sets

The first set of histograms (Figure 4.18), shows the percentage of trusted gene sets related to cancer (3,079 gene sets). Some centralities, for both cancers, are capable to capture gene sets enriched in our ranked list. We can highlight that Protein networks can capture more gene sets than RNA layer. We believe this result is a consequence of RNA network being more sparse (disconnected) than Protein.

We quantify the *na_pos* and *na_neg*, that are gene sets found in the extremes of our ranked list (Figure 4.18). To show the difference in quantity, we separate each bar in two shades of transparency. The most transparent, top of the bar, represents the gene sets enriched at top list (*na_pos*), and the least transparent, bottom of the bar, represents the gene sets enriched at the bottom (*na_neg*).

The enrichment analysis show that we find more gene sets in *na_neg* than in *na_pos* in most networks. But at Breast RNA Network, we see centralities that found gene sets in the top list and not at bottom. One hypothesis is the sparseness of Breast RNA network. Not only we have less connected nodes through correlations but we also have disconnected components that we take into account.

For each centrality, we can see some particularities: (1) *betweenness* is not monotonic through correlations and network types, the disconnected components imply this, but we highlight the *na_pos* in Protein networks; (2) *closeness* follows a uniform behavior and captures more gene sets as we increase the correlation, we highlight the Breast Protein type where this centrality captured more gene sets; (3) *clustering* is not monotonic through correlations, but is able to compete in amount of gene sets and we highlight the Ovarian RNA where it captures more gene sets than other centralities; (4) *diffusion* is monotonic and we highlight the uniformity in Ovarian RNA where it captures more than 7.5% of gene sets; (5) *dmnc* looks not uniform in some networks through correlations but capture more *na_pos* than other centralities through the network types; (6) *eigenvector* is monotonic through correlations thresholds and we highlight Breast RNA and Ovarian Protein types where this centrality captured more gene sets; (7) *laplacian* looks more monotonic in Ovarian networks; (8) *leverage* is not uniform but able to capture more *na_pos* than other centralities in Ovarian Protein networks; (9) *topological* is not monotonic and unable to compete against other centralities in amount of gene sets; (10) *weighted* is uniform and is able to compete in quantity of gene sets.



(a) Breast Protein

(b) Breast RNA

(c) Ovarian Protein

(d) Ovarian RNA

Figure 4.18: Percentage of Cancer Only Histogram for MSigDB of Enriched Gene Sets with FDR < 0.1.

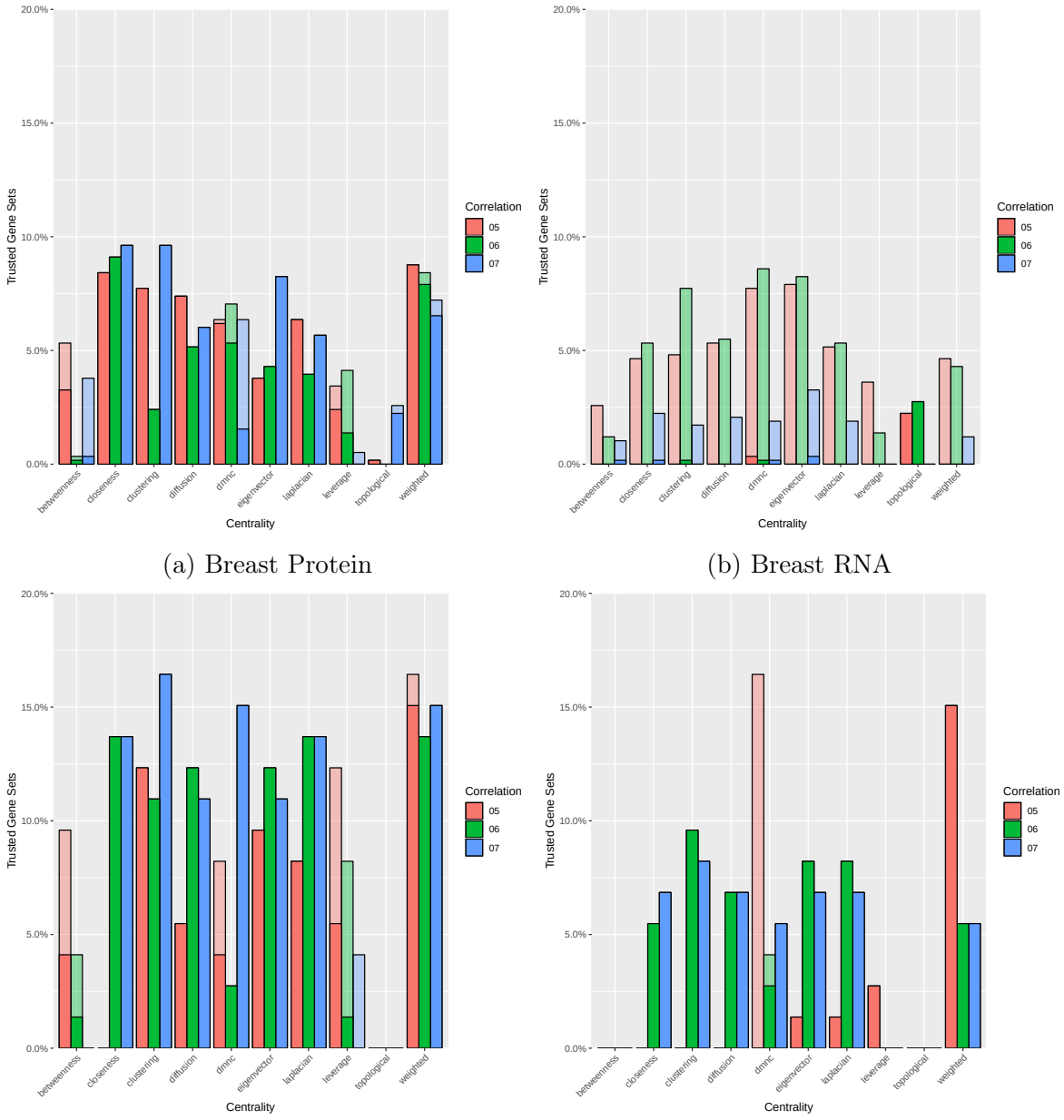
Cancer Type Only (Breast/Ovarian) Gene Sets

When we take a look on each cancer type, based on our labels, we have 582 Breast Cancer gene sets and 73 Ovarian Cancer gene sets. Figure 4.19 shows percentage histograms for each cancer type. Here we can also show that Protein Layer has better percentage results. Additionally we highlight that in Breast Protein our centralities can find almost 10% of all Breast Cancer related gene sets. This result could also be a consequence of RNA network being more sparse (disconnected) than Protein.

We applied same analysis as we did for all cancer gene sets, as shown in Figure 4.19, quantifying the *na_pos* and *na_neg*, that are gene sets found in the extremes of our ranked list. We separate each bar in two shades of transparency following as: the most transparent, top of the bar, represents the gene sets enriched at top list (*na_pos*), and the least transparent, bottom of the bar, represents the gene sets enriched at the bottom (*na_neg*).

Most of the enrichment analysis show that we find more gene sets in *na_neg* than in *na_pos*. The bottom of our ranked lists seem more important than the top when we quantify the gene sets under FDR 0.1. Another analysis that we can look is the quantity of Ovarian Cancer Gene Sets found in correlation 0.5 in RNA networks (Figure 4.19d). Most centralities can barely capture a single gene set, but the *dmnc* and *weighted* centralities are able to capture, and what is more interesting is that they get the highest amount of gene sets with 0.5.

For each centrality, we can see some particularities: (1) *betweenness* is not monotonic through correlations and network types, and we highlight the *na_pos* in Protein networks; (2) *closeness* follows a uniform behavior, we highlight the Breast Protein type where this centrality captured more gene sets; (3) *clustering* is not monotonic through correlations, but is able to compete in amount of gene sets and we highlight the Breast Protein where it captures more gene sets; (4) *diffusion* is monotonic and we highlight the uniformity in Ovarian RNA where it capture more than 7.5% of gene sets; (5) *dmnc* looks not uniform in some networks through correlations but capture more *na_pos* in than other centralities through the network types, we highlight this centrality in RNA networks where it captures more gene sets; (6) *eigenvector* is not monotonic through correlations; (7) *laplacian* is monotonic in Ovarian networks; (8) *leverage* is not uniform but able to capture more *na_pos* than other centralities in Ovarian Protein networks; (9) *topological* is not monotonic and unable to compete against other centralities in amount of gene sets; (10) *weighted* is uniform and is able to compete in quantity of gene sets, we highlight Ovarian Protein where it gets more gene sets.



(a) Breast Protein

(b) Breast RNA

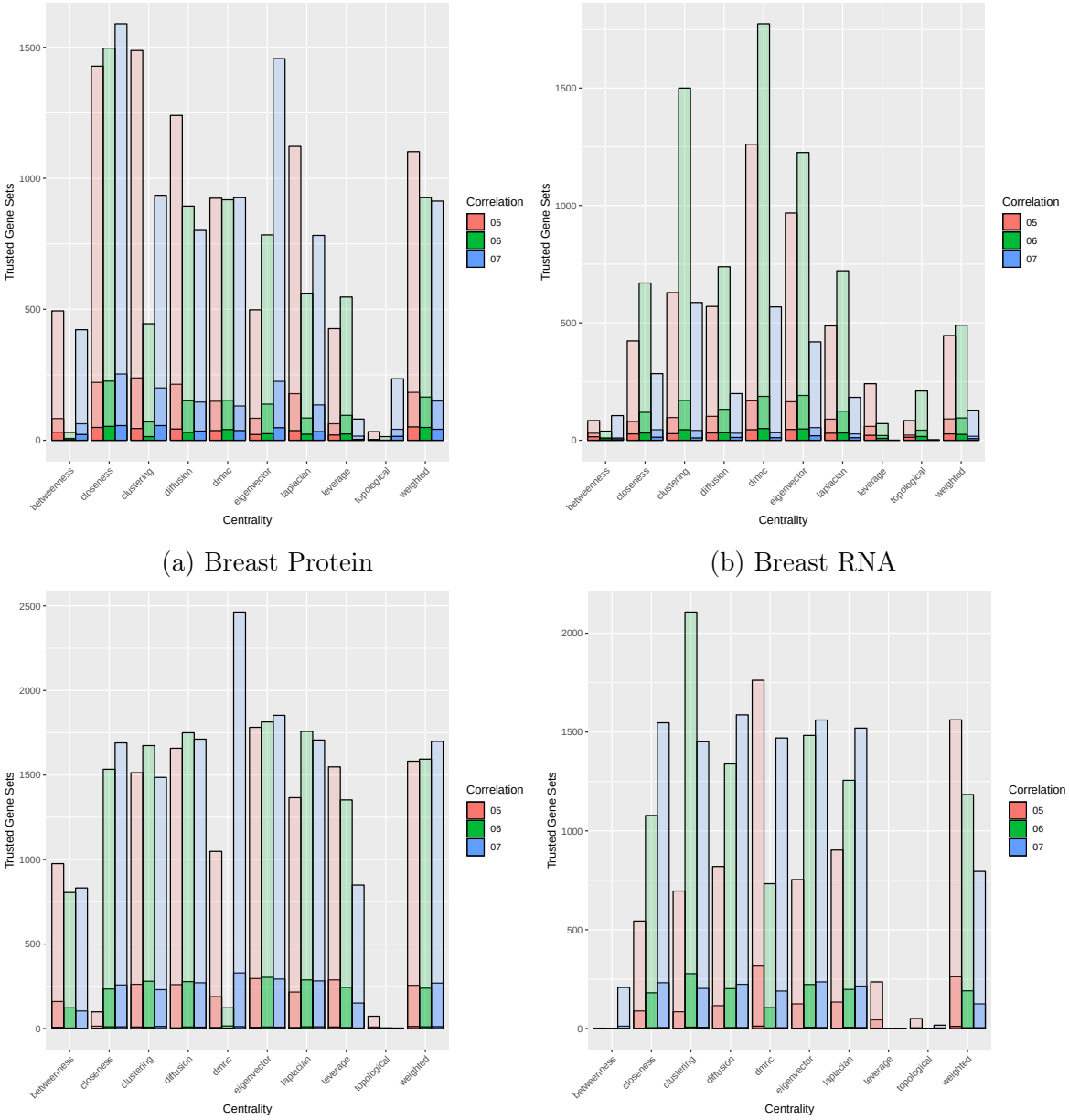
(c) Ovarian Protein

(d) Ovarian RNA

Figure 4.19: Percentage of Cancer Type Only Histogram for MSigDB of Enriched Gene Sets with FDR < 0.1.

All Gene Sets from MSigDB

In Figure 4.20, we show absolute quantity of Trusted Gene sets, where we separate each bar in three shades of transparency. From the most transparent to less, we divide in: not cancer, cancer, and cancer type (Breast and Ovarian) gene sets.



(a) Breast Protein

(b) Breast RNA

(c) Ovarian Protein

(d) Ovarian RNA

Figure 4.20: Histogram for MSigDB of Enriched Gene Sets with FDR < 0.1.

This histograms show a comparison between not cancer gene sets found with cancer gene sets and cancer type gene sets. This amount of gene sets found could be a result of, not only cancer analysis, but also biological processes involved in our network. We believe this "not cancer" related gene sets could be further explored.

4.4 Chapter Discussion

In this chapter, we describe experiments to capture biological processes and gene sets for Breast and Ovarian cancer in our approach. For this reason, we proposed two types of analysis: multi layer analysis, evaluating the groups in all layers (CNA, RNA and Protein); and single layer analysis, evaluating ranked lists in networks of RNA and Protein layers.

In our analysis for multi layer data, we applied Community Detection to separate groups of gene expression data. We found out that the largest community of genes and proteins shared by both cancers (Breast and Ovarian) is a group biologically related to metabolic and cellular processes. We concluded that our network science approach is mirroring the behavior of a cell.

For single layer data, we compared different centrality measures to produce inputs for enrichment analysis. As a result, we could see that centralities can capture trusted gene sets. The percentage histograms show that protein layer have more consistency results through our centralities. Each centrality have individual behaviours through our network types and correlations.

5

Conclusions

In this work, we proposed approaches of Network Science to characterize Breast and Ovarian cancer evaluated through enrichment analysis. The challenge for this work was the low number of samples, which made us avoid machine learning based representation and select a network based approach.

In our approach, we explore network characteristics as inputs for Enrichment Analysis. After we decided configurations as correlation thresholds, enrichment analysis methods, centralities and communitiy detection methods, our work is ready to run with any gene expression data.

5.1 Final Remarks

In multi layer data we could reply existing biological process. We could find gene lists common in Breast and Ovarian cancer. Our community detection, unsupervised, could find that the biggest group, present in both cancers, is biologically related to metabolic and cellular process.

For single omic data, we applied centralities capable of producing inputs for enrichment analysis. As a result, we could see not only that centralities are able to capture trusted gene sets, but also that Protein layer shows more consistency to capture those gene sets. To set different thresholds lead to more interesting results.

Finally, we identified gene sets, in a collection of studied gene sets, related to our lists. This finding show us that extracting characteristics using Network Science for Enrichment Analysis lead us to biological characteristics within Cancer type. Those gene sets found could define characteristics for our specific type of cancer.

5.2 Limitations

Working with the absolute value of the correlation coefficient for our edges can hide some characteristics for low expressed genes and proteins. But as we said before, this positive edges are necessary to work with some algorithms both for community detection and centrality measures.

We did not experiment for all centrality measures, we chose only neighborhood and distance based centralities, our hypothesis was that genes and proteins have pathways, leading us to distance based, and they also affect their surroundings, neighborhood based. We did not explore all communities in multi layer experiment because we only focused on the largest community from both cancers, maybe there are small communities not responsible for expressive biological processes but they are captured in our detection.

We did not evaluate healthy samples, it could help us to filter gene sets that are related to normal cell process. Cancer networks do not represents only cancer gene sets, we believe that healthy samples could lead us to a better understanding of what is only cancer related.

5.3 Future Works

For future works, we could use our approach of characterization through network topology to explore ways for characterize Ovarian Cancer subtypes. Breast cancer has its subtypes studied and Ovarian cancer has some way to go for exploring his subtypes. As we saw in our first experiment, even with heterogeneity of Breast and Ovarian cancer, we are able to see intersections between their communities in multi omic. That is, even with their diversity we were able to see some characteristics in both.

As a new approach of Ovarian Cancer subtypes, we could apply our knowledge in network topology characterization and cluster samples in Ovarian that shares same behavior as Breast samples labeled with their subtypes. After that, we could reconstruct networks for each cluster of Ovarian to extract their characteristics following our approach in this work, which we could embed as characteristics for these new subtypes.

Bibliography

Assenov, Y., Ramírez, F., Schelhorn, S.-E., Lengauer, T., and Albrecht, M. (2007). Computing topological parameters of biological networks. *Bioinformatics*, 24(2):282–284.

Atay, Y., Koc, I., Babaoglu, I., and Kodaz, H. (2017). Community detection from biological and social networks: A comparative analysis of metaheuristic algorithms. *Applied Soft Computing*, 50:194–211.

Bastian, M., Heymann, S., and Jacomy, M. (2009). Gephi: an open source software for exploring and manipulating networks. In *Third international AAAI conference on Weblogs and Social Media*.

Bedi, P. and Sharma, C. (2016). Community detection in social networks. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 6(3):115–135.

Benesty, J., Chen, J., Huang, Y., and Cohen, I. (2009). Pearson correlation coefficient. In *Noise Reduction in Speech Processing*, pages 1–4.

Beveridge, A. and Shan, J. (2016). Network of thrones. *Math Horizons*, 23(4):18–22.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and experiment*, 2008(10):P10008.

Borgatti, S. P. (1995). Centrality and aids. *Connections*, 18(1):112–114.

Chang, H.-H., Dreyfuss, J. M., and Ramoni, M. F. (2011). A transcriptional network signature characterizes lung cancer subtypes. *Cancer*, 117(2):353–360.

Chang, H.-H. and Ramoni, M. F. (2009). Transcriptional network classifiers.

Chasman, D., Siahpirani, A. F., and Roy, S. (2016). Network-based approaches for analysis of complex biological systems. *Current Opinion in Biotechnology*, 39(1):157–166.

Cheadle, C., Vawter, M. P., Freed, W. J., and Becker, K. G. (2003). Analysis of microarray data using z score transformation. *The Journal of Molecular Diagnostics*, 5(2):73–81.

Chen, D.-B., Gao, H., Lü, L., and Zhou, T. (2013). Identifying influential nodes in large-scale directed networks: the role of clustering. *PloS One*, 8(10):e77455.

Consortium, E. P. et al. (2012). An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414):57–74.

Crick, F. (1970). Central dogma of molecular biology. *Nature*, 227(5258):561–563.

D'haeseleer, P. (2005). How does gene expression clustering work? *Nature Biotechnology*, 23(12):1499–1501.

Dutta, B., Pusztai, L., Qi, Y., André, F., Lazar, V., Bianchini, G., Ueno, N., Agarwal, R., Wang, B., Shiang, C. Y., et al. (2012). A network-based, integrative study to identify core biological pathways that drive breast cancer clinical subtypes. *British Journal of Cancer*, 106(6):1107–1116.

Fessler, E., Jansen, M., Melo, F. D. S. E., Zhao, L., Prasetyanti, P., Rodermond, H., Kandimalla, R., Linnekamp, J., Franitza, M., van Hooff, S., et al. (2016). A multidimensional network approach reveals micrornas as determinants of the mesenchymal colorectal cancer subtype. *Oncogene*, 35(46):6026–6037.

Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, 486(3-5):75–174.

Girvan, M. and Newman, M. E. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826.

Goldhirsch, A. ., Wood, W., Coates, A., Gelber, R., Thürlimann, B., Senn, H.-J., and members, P. (2011). Strategies for subtypes—dealing with the diversity of breast cancer: highlights of the st gallen international expert consensus on the primary therapy of early breast cancer 2011. *Annals of oncology*, 22(8):1736–1747.

Graudenzi, A., Cava, C., Bertoli, G., Fromm, B., Flatmark, K., Mauri, G., and Castiglioni, I. (2017). Pathway-based classification of breast cancer subtypes. *Frontiers in Bioscience*, Landmark Edition(22):1697–1712.

Griffiths, T. L., Steyvers, M., and Firl, A. (2007). Google and the mind: Predicting fluency with pagerank. *Psychological Science*, 18(12):1069–1076.

Gu, Z., Gu, L., Eils, R., Schlesner, M., and Brors, B. (2014). circlize implements and enhances circular visualization in r. *Bioinformatics*, 30(19):2811–2812.

Hua, L., Li, L., and Zhou, P. (2013). Identifying breast cancer subtype related mirnas from two constructed mirnas interaction networks in silico method. *BioMed Research International*, 2013.

Joyce, K. E., Laurienti, P. J., Burdette, J. H., and Hayasaka, S. (2010). A new measure of centrality for brain networks. *PLoS One*, 5(8):e12200.

Kernighan, B. W. and Lin, S. (1970). An efficient heuristic procedure for partitioning graphs. *The Bell System Technical Journal*, 49(2):291–307.

Khatri, P., Sirota, M., and Butte, A. J. (2012). Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS computational biology*, 8(2):e1002375.

Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M. J., et al. (2015). Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–330.

Langfelder, P. and Horvath, S. (2008). Wgcna: an r package for weighted correlation network analysis. *BMC Bioinformatics*, 9(1):559–571.

Leydesdorff, L. (2007). Betweenness centrality as an indicator of the interdisciplinarity of scientific journals. *Journal of the American Society for Information Science and Technology*, 58(9):1303–1319.

Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H., Tamayo, P., and Mesirov, J. P. (2011). Molecular signatures database (msigdb) 3.0. *Bioinformatics*, 27(12):1739–1740.

Lin, C.-Y., Chin, C.-H., Wu, H.-H., Chen, S.-H., Ho, C.-W., and Ko, M.-T. (2008). Hubba: hub objects analyzer—a framework of interactome hubs identification for network biology. *Nucleic Acids Research*, 36(2):438–443.

Liu, Z., Zhang, X.-S., and Zhang, S. (2014). Breast tumor subgroups reveal diverse clinical prognostic power. *Scientific Reports*, 4(4002).

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., et al. (2013). The genotype-tissue expression (gtex) project. *Nature Genetics*, 45(6):580.

Mertins, P., Mani, D., Ruggles, K. V., Gillette, M. A., Clauser, K. R., Wang, P., Wang, X., Qiao, J. W., Cao, S., Petralia, F., et al. (2016). Proteogenomics connects somatic mutations to signaling in breast cancer. *Nature*, 534(7605):55.

Mi, H., Muruganujan, A., Huang, X., Ebert, D., Mills, C., Guo, X., and Thomas, P. D. (2019). Protocol update for large-scale genome and gene function analysis with the panther classification system (v. 14.0). *Nature Protocols*, 14(1):703–721.

Mirnezami, R., Nicholson, J., and Darzi, A. (2012). Preparing for precision medicine. *New England Journal of Medicine*, 366(6):489–491.

Mukaka, M. M. (2012). A guide to appropriate use of correlation coefficient in medical research. *Malawi Medical Journal*, 24(3):69–71.

Newman, M. E. and Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113.

Okamoto, K., Chen, W., and Li, X.-Y. (2008). Ranking of closeness centrality for large-scale social networks. In *International Workshop on Frontiers in Algorithmics*, volume 5059, pages 186–195.

Ozols, R. (2005). Treatment goals in ovarian cancer. *International Journal of Gynecological Cancer*, 15(1):3–11.

Pal, S. K., Kundu, S., and Murthy, C. (2014). Centrality measures, upper bound, and influence maximization in large scale directed social networks. *Fundamenta Informaticae*, 130(3):317–342.

Pecorelli, S., Favalli, G., Zigliani, L., and Odicino, F. (2003). Cancer in women. *International Journal of Gynecology & Obstetrics*, 82(3):369–379.

Perou, C. M., Jeffrey, S. S., Van De Rijn, M., Rees, C. A., Eisen, M. B., Ross, D. T., Pergamenschikov, A., Williams, C. F., Zhu, S. X., Lee, J. C., et al. (1999). Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proceedings of the National Academy of Sciences*, 96(16):9212–9217.

Perou, C. M., Sorlie, T., Eisen, M. B., Van De Rijn, M., et al. (2000). Molecular portraits of human breast tumours. *Nature*, 406(6797):747–752.

Pizzuti, C. (2008). Ga-net: A genetic algorithm for community detection in social networks. In *International Conference on Parallel Problem Solving from Nature*, pages 1081–1090. Springer.

Qi, X., Fuller, E., Wu, Q., Wu, Y., and Zhang, C.-Q. (2012). Laplacian centrality: A new centrality measure for weighted networks. *Information Sciences*, 194:240–253.

Raghavan, U. N., Albert, R., and Kumara, S. (2007). Near linear time algorithm to detect community structures in large-scale networks. *Physical Review E*, 76(3):036106.

Raymond, M. and Rousset, F. (1995). An exact test for population differentiation. *Evolution*, 49(6):1280–1283.

Subramanian, A., Kuehn, H., Gould, J., Tamayo, P., and Mesirov, J. P. (2007). Gsea-p: a desktop application for gene set enrichment analysis. *Bioinformatics*, 23(23):3251–3253.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550.

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth, A., Santos, A., Tsafou, K. P., et al. (2014). String v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic acids research*, 43(D1):D447–D452.

Tang, X., Wang, J., Zhong, J., and Pan, Y. (2014). Predicting essential proteins based on weighted degree centrality. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 11(2):407–418.

Van't Veer, L. J., Dai, H., Van De Vijver, M. J., He, Y. D., Hart, A. A., Mao, M., Peterse, H. L., Van Der Kooy, K., Marton, M. J., Witteveen, A. T., et al. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415(6871):530–536.

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C., Stuart, J. M., Network, C. G. A. R., et al. (2013). The cancer genome atlas pan-cancer analysis project. *Nature Genetics*, 45(10):1113–1120.

Yang, L., Wang, S., Zhou, M., Chen, X., Jiang, W., Zuo, Y., and Lv, Y. (2017). Molecular classification of prostate adenocarcinoma by the integrated somatic mutation profiles and molecular network. *Scientific Reports*, 7(1):738–751.

Yersal, O. and Barutca, S. (2014). Biological subtypes of breast cancer: Prognostic and therapeutic implications. *World Journal of Clinical Oncology*, 5(3):412–424.

Zhang, D., Chen, P., Zheng, C.-H., and Xia, J. (2016). Identification of ovarian cancer subtype-specific network modules and candidate drivers through an integrative genomics approach. *Oncotarget*, 7(4):4298–4309.