2017

# Identifying Network-Biomarkers of Breast Cancer Survivability

Sheikh Abdullah Al Jubair
*University of Windsor*

# Identifying Network-Biomarkers of Breast Cancer Survivability

By

**Sheikh Abdullah Al Jubair**

A Thesis
Submitted to the Faculty of Graduate Studies
through the School of Computer Science
in Partial Fulfillment of the Requirements for
the Degree of Master of Science
at the University of Windsor

Windsor, Ontario, Canada

2017

Identifying Network-Biomarkers of Breast Cancer Survivability

by

Sheikh Abdullah Al Jubair

APPROVED BY:

L. Porter
Department of Biological Sciences

R. Gras
School of Computer Science

A. Ngom, Advisor
School of Computer Science

L. Rueda, co-Advisor
School of Computer Science

May 1, 2017

# DECLARATION OF ORIGINALITY

I hereby certify that I am the sole author of this thesis and that no part of this thesis has been published or submitted for publication.

I certify that, to the best of my knowledge, my thesis does not infringe upon anyoneâĂŹs copyright nor violate any proprietary rights and that any ideas, techniques, quotations, or any other material from the work of other people included in my thesis, published or otherwise, are fully acknowledged in accordance with the standard referencing practices. Furthermore, to the extent that I have included copyrighted material that surpasses the bounds of fair dealing within the meaning of the Canada Copyright Act, I certify that I have obtained a written permission from the copyright owner(s) to include such material(s) in my thesis and have included copies of such copyright clearances to my appendix.

I declare that this is a true copy of my thesis, including any final revisions, as approved by my thesis committee and the Graduate Studies office, and that this thesis has not been submitted for a higher degree to any other University or Institution.

# ABSTRACT

One of the key challenges of breast cancer research is to predict whether a patient identified with specific subtype or treated with a specific therapy is going to survive or die. Current studies find small subsets of gene biomarkers able to accurately predict the response to therapy. In these studies, the selected genes are not necessarily functionally related, and hence, they may not correctly indicate the molecular mechanism behind breast cancer survivability. Also, several studies have shown there is a very low overlap between the different respective biomarkers subsets for the same cancer disease. To improve the robustness of classification performance and stability of detected biomarkers, recent methods take existing knowledge on relations between genes into account in the classifier, by aggregating functionality related genes to produce discriminative gene subnetworks called network-biomarkers. In this paper, given a breast cancer dataset of patients with different subtypes treated with a given therapy drug, we devised network-based machine learning approach by integrating protein protein interaction network (PPI) with gene expression data (1) to identify the network-biomarkers of breast cancer survivability a) based on subtypes and b) based on therapy and (2) to predict the survivability of breast cancer patients a) based on subtypes b) treated with a therapy drug. We used the concept of seed gene for identification of network-biomarkers with distance 2, 3 and 4 from seed gene protein and our method found distance 3 and 4 are the distance that gives us best result for identifying survivability of breast cancer patient based on subtype and therapy respectively. To solve the class imbalance problem in some subtypes, we implemented ADASYN. We obtained best classification performance using random forest where the geometric mean, F1-measure and accuracy are respectively 0.867, 0.850 and 87.00% for subtype specific study, and 0.829, 0.807 and 83.77%, for therapy specific.

# DEDICATION

I would like to dedicate my thesis to my dear brother, father, and specially to my mother who always persuaded me to take this program and encourage me in difficult situations.

# AKNOWLEDGEMENTS

I would like to express my deepest appreciation to my supervisor Dr. Alioune Ngom, and my co-supervisor Dr. Luis Rueda for their perpetual supports, patience, and inspiration during my Master's program in University of Windsor. It was such an honor for me to know you, and also be in your research team. Thank you so much for giving me an opportunity to learn from you.

Secondly, I would also like to express my gratitude to my committee members Dr. Lisa Porter, and Dr. Robin Gras for their beneficial advices and suggestions to my thesis.

Meanwhile I would like to express my special thanks to my friends for helping me during past two years.

Finally, I would like to express my greatest appreciation to my family for all unconditional love, support, patience, encouragement, and kindness they gave me during my whole life.

TABLE OF CONTENTS

## LIST OF TABLES

## LIST OF FIGURES

# CHAPTER 1

# *Introduction*

## 1.1 Breast Cancer and Important Terminologies

Breast Cancer is a complex disease that starts in the cells of the breast. It comprises five [23] to ten [7] genomics subtypes where each subtypes has distinct molecular mechanism and has distinct clinical progression. Recent studies showed that there is extensive diversity between and within breast cancer patients and each breast cancer shows unique characteristics. The heterogeneity of this cancer complicates diagnosis and treatment as studies found many patient of breast cancer undergoes with over-treatment [29]. The reason behind is thec mutation of small number of genes called drivers whose change deregulate many biologial processes or cellular pathways, and therefore leading to initiation and progression of breast cancer as well as resistance to treatment [16]. Driver genes are expressed aberrantly that confers selective growth or contain driver gene mutation [33]. Driver genes also may contain passenger gene mutation that have no direct or indirect effect on selective growth advantage of the cell [35]. Passenger genes are those whose deregulations or expression changes are the by product of drivers. Thus the driver genes are those that are responsible for cancer and the passengers have no effect on cancer but they can be present in the driver gene mutation. Together, the drivers and their passengers are called gene biomarkers. For predicting breast cancer classes, it is necessary to discover the biomarkers that play vital role in breast cancer subtypes. There are 8 different kinds of biomarkers: 1) prognostic biomarkers: to predict the development of a cancer [14], 2) diagnostic biomarkers: to predict the presence of disease or condition of interest or the subtype of cancer [22], 3) predictive biomarkers: to predict the survivability of patient treated with specific drug[8], 4) treatment biomarkers: to predict the effectiveness of a treatment[24]. 5) progression biomarkers: to predict whether the cancer is spreading or not [30], 6) monitoring biomarkers: to predict if therapy is working [25], 7) recurrence biomarkers: to predict whether the cancer will recur after sometimes or not

[19], 8) risk biomarkers: to predict predisposition to cancer [32]. In this thesis, we focus on finding predictive biomarkers that can identify survival of breast cancer patient.

Gene expression METABRIC dataset has been used to identify differentially expressed genes that can predict the survival of breast cancer patient. The word METABRIC comes by taking initial letters from Molecular Taxonomy of Breast Cancer International Consortium. It is a Canada-UK project to classify breast cancer into subcategories based on their molecular signatures so that it will be helpful for identifying optimal treatment of breast cancer patient. Gene expression is the process by which the instruction in our DNA are converted into a functional product, such as protein. A classifier can predict breast cancer class from an unseen data based on previously trained data of known classes. The fundamental limitation of gene expression base study is that it fails to obtain robust and highly predictive classifier across different datasets as C Soneson et al. [28] found the classifier's accuracy decreases dramatically when the classifier trained with the dataset of one study and test with the dataset of another study for the same subtype. It is also found that even for the same breast cancer subtype only a 4% gene biomarkers overlaps across different datasets[10]. This leads to the hypothesis that adding more information with gene expression data with the genes that are functionally related can predict the breast cancer classes and give robustness of classifier performance to the prediction model. Network-biomarkers takes the existing knowledge on relation among genes into account by aggregating functionally related genes to produce discriminative gene subnetworks. Studies showed that network-biomarkers significantly improves the classifier performance for different datasets and produce stable network-biomarkers [2] [3] [6] [8] [36]. In this thesis, we devised a machine learning approach that used protein protein interaction network along with gene expression data to find the functional relation among genes and their corresponding protein for predicting the survivability of breast cancer patient and for identifying network-biomakers for survivability.

There are three types of therapy: chemotherapy, hormone therapy and radiotherapy. A breast cancer patient can receive any combination of these three therapy. Even the patient may not receive any therapy too. Chemotherapy is a medication or combination of medication that is used to treat cancer patient. On the other hand hormone therapy is used only for those patients who have hormone receptor positive breast cancer. The purpose of hormone therapy is to reduce the amount of hormone estrogen in the body and to block the action of estrogen in the body. Radiotherapy uses high energy ray to damage cancer cells and can be used externally or internally. Survival of patients are dependent on choosing appropriate therapy and identifying appropriate biomarkers. Patient survival also depends on subtypes as studies found that some subtypes have a higher mortality than other subtypes [5].

In Chapter 2 we introduce our method. In Chapter 3, we talk about our experiment and results. In Chapter 4, we discuss about our method and its limitation and Chapter 5 is our conclusion.

## 1.2 Problem Statement

In this thesis, given a breast cancer dataset of patients of different subtypes and treated with different combination of chemotherapy, hormone therapy and radiotherapy , our aim is (1) to identify the network-biomarkers of breast cancer survivability a) based on subtypes without considering any therapy b) based on different combination of therapy but not considering the subtypes and (2) to predict the survivability of breast cancer patients a) of different subtypes (b) treated with different combinations of chemotherapy, hormone therapy and radio therapy or no drug.

## 1.3 Motivation

Current bioinformatics methods find small subsets of gene biomarkers able to accurately predict the response to therapy. In these studies, the selected genes are not necessarily functionally related, and hence, they may not correctly indicate the molecular mechanism behind breast cancer survivability. Several studies have shown the lack of robustness and stability of currently identified biomarker subsets from different research; that is, there is a very low overlap between the different respective biomarkers subsets for the same cancer disease, and hence a subset identified in one study yields low performance on data used in another study. This lack of robustness and stability is because of insufficient patient sample size [18] that is the number of features are generally too high compared to number of samples and generally, there are very few samples available for cancer study, the inherent measurement noise in microarray experiments [10] that is the subsets of genes that are selected by feature selection technique varies across different subset of patients for the same disease, insufficient knowledge capture in tumor samples or the heterogeneity in tumor samples [29]. Computational methods based on detecting differentially expressed genes do not consider the dependencies or relationships between genes to accurately classify the sample data; thus, identified gene biomarker sets may not contain driver genes or may contain many differentially expressed genes with redundant information yielding decreased prediction performance [21].

Network based methods that takes the knowledge of relation among genes into account and generates more stable network-biomarkers than the gene biomarkers that can accurately predict the outcome of breast cancer patient across different dataset [6] [2] [3] [8] [9] [36]. As studies found that different protein and gene interaction plays an important role in cancer molecular mechanism, network-biomarkers with one or more secondary network with gene expression data gives more information for survivability of breast cancer patient. These motivates us to find network-biomarkers than gene biomarkers to predict the survivability of patients of different subtypes and treated with different combination of drugs.

## 1.4 Literature review

Network-biomarkers identification problem is a subset selection problem in which the selected genes in a subset are functionally related. The best network-biomarkers are those that discriminate the two classes of breast cancer and in our case the survival of patients which means whether the breast cancer patient is going to survive or not. Most of the network based approach starts with a seed gene as initial subnetwork and expand iteratively from seed gene to its neighboring gene. The neighbor gene is generally added to the subnetwork based on improvement on some discriminative score of subnetwork with the newly added gene.

In 2007, Chuang et al. [6] takes the knowledge of protein protein interaction network with gene expression data to identify the network markers and predict the breast cancer metastasis for patient. His identified subnetworks were actually meta-genes where each meta-gene is a subnetwork in PPI and he showed that network-biomarkers increase the classifiers accuracy between two datasets compared to gene expression biomarkers.

In 2011, P. Dao et al. proposed OptDis [8] a network based classification algorithm that used color coding technique along with protein protein interaction network. They used OptDis for predicting response to drug and found it provides better and more stable performance than other network based and gene based approach.

EB. van den Akker et al. [2] integrated protein protein interaction network along with gene gene co expression network to predict breast cancer metastasis. For this study, they used pair wise gene expression correlation measures to identify seed genes. To add the information of PPI network and gene gene co-expression network, they applied different integration approaches. Their method obtained the robustness of classifier performance as their network-biomarkers works well across 6 different datasets.

K. Zhang et al. [37] devised a new method called CAERUS that integrates the information of protein structure, protein-protein interaction networks and gene expression data. They developed a scoring mechanism for each protein by considering domain connections to its interacting partner and somatic mutation present in the domain. Then they identified those gene signatures that have the score value above a predefined threshold. Then they calculated the correlation of gene expression and its protein, and used Naive Bayes classifier to predict the breast cancer carcinoma.This is the first predictive method to classify cancer outcomes based on the relationship between domain organization and protein network.

Dutta et al. [9] identified network-biomarkers for predicting ER+, Her2+ and triple negative breast cancer for patients. In their network they integrated the information of gene expression and gene copy number data on protein-protein interaction, transcriptional-regulatory and signalling networks. They also showed that their network based approach is reproducible and functionally important.

D Wu et al. [36] showed that classifiers with knowledge on relation between genes into account, improve robustness of classification performance and stability of detected biomarkers. In 2014 Garcia et. al [11] proposed an interactome based approach that takes copy number variation and protein protein interaction network into account and predicts relapse free survival of breast cancer patient.

In 2015 A. Allahyar et al. [3] proposed FERAL. FERAL is based on sparse group lasso that selects the genes for network during training. It uses the concept of metagenes by summarizing multiple genes with different operators where each metagene is a subnetwork.

There are also many network based methods to predict the breast cancer classes.

## 1.5 Contribution

In this thesis, we propose a novel approach to identify network-biomarkers and predict survivability of breast cancer patients. Our method is based on:

- chi square feature selection technique to identify seed genes
- subnetwork identification from PPI using Chuang et al's approach
- chi square and forward feature selection with random forest algorithm to identify metagenes that discriminate our classes.
- ADASYN to solve class imbalance within the data.
- random forest as classifier to predict survivability of breast cancer patients.

In this chapter, we discuss about some important terminology regarding breast cancer, what problem we want to solve in our thesis, what motivates us to work with this problem and their related literature.

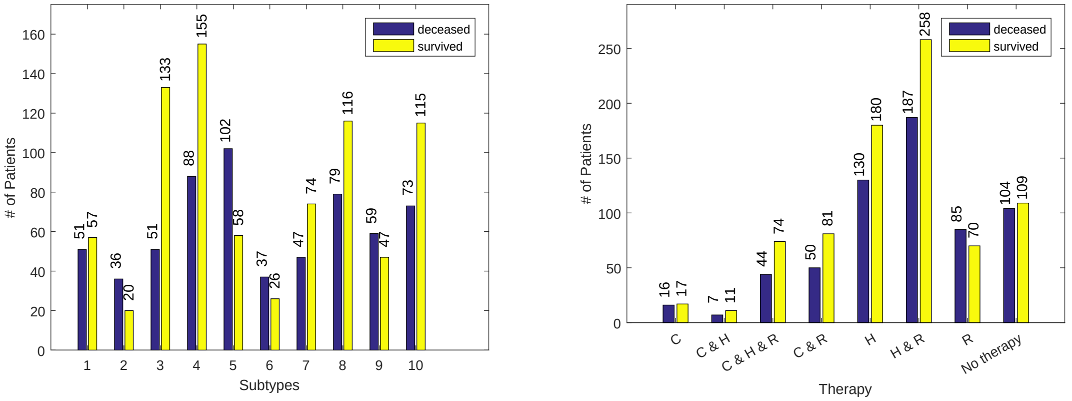
# CHAPTER 2

# *Materials and Methods*

## 2.1 Dataset

We used the original METABRIC dataset [7] publicly available in cBioPortal [1]. The dataset contains gene expression, copy number and somatic mutation data along with clinical data (tumour morphology, ER and HER2 status, patient characteristics, treatment, follow-data, survival status data) for 1904 patients. Among these patients, 480 of them died for other reasons and we remove them from dataset that gives us 1,424 patients. For predicting the survivability of the patients based on subtypes, we separate the data according to ten different subtypes. As we have three types of therapy in the dataset that are chemotherapy, hormone therapy and radio therapy, we obtain eight combination of these therapies including a patient did not receive any therapy. For predicting the survivability of the patient treated with any combination of these drug, we separate the dataset in eight subsets based on treatment they received. Figure 2.1 shows the number of samples in each subtype and in each combination of drug. The copy number aberration and copy number variations were generated using Affymetrix SNP 6.0 arrays and gene expression data were obtained using Illumina HT 12 technology. The dataset has 24,368 differentially expressed genes.

Our PPI network contains 224,766 interactions with proteins. The protein protein interaction are gathered from a number of sources: BioGrid, Mint, InnateDB, PDB. We have in total of 15,123 unique proteins and 14,845 genes that are mapped to one or more proteins. We have 12, 962 genes that are common in our gene expression dataset and in PPI network.



(a) Number of patients in each subtypes.

(b) Number of patients in each combination of therapy.

FIGURE 2.1: Subtype and therapy specific samples.

## 2.2 Chi square

Chi square feature selection technique measures the degree of independence of each feature classes. It uses the following formula:

$$\chi^2(Y, X) = \frac{N \times (AD - CB)}{(A + C) \times (B + D) \times (A + B) \times (C + D)} \tag{2.1}$$

A is the number of times feature $X$ in our case discretized gene expression $X$ and class $Y$ that is survived or deceased co-occur. $B$ is the number of times X occurs without Y. C is the number of times Y occurs without X. D is the number of times neither X and Y occurs. N is the total number of samples.

## 2.3 Obtaining Subnetwork from PPI

In our thesis, a subnetwork is a single connected component in protein protein interaction network. For obtaining subnetworks from seed genes we implemented Chuang et. al [6] approach.

Algorithm 1 shows how to obtain a subnetwork from PPI using a seed gene.Subnetwork identification from seed gene is a greedy approach. Given a seed gene, we expand the network with $distance = 1$ from seed gene protein and consider each child protein with seed gene protein as separate network. We then calculate the activity score of each of the

```
Input: Seed Gene, PPI, Gene Expression
Result: Subnetwork
Given a seed gene find corresponding proteins and add them to seed proteins list;
for each seed proteins do
    initialize parent protein with seed protein;
    initialize subnetwork with parent protein;
    find all proteins with d = 1 from parent protein;
    add them to other protein list;
    for each other protein do
        activity score of sample j, $a_j = \sum_{i=1}^{n} \frac{z_{ij}}{\sqrt{n}}$;
        discretize activity score, $a'$;
        Mutual information, $MI(a', c) = \sum_{x \epsilon a'} \sum_{y \epsilon c} p(x,y) log \frac{p(x,y)}{p(x).p(y)}$
    end
    add protein that gives highest mutual information to subnetwork list;
    add parent of the protein to parent protein list;
    remove the protein from other protein list;
    if distance from seed protein to recently added protein <maxdistance then
        expand network with d = 1 from the added protin;
        add these proteins to other protein list;
    end
    if No improvement in Mutual Information or Mutual Information >0.5 then
        break;
    end
end
```

**Algorithm 1:** Subnetwork identification with a seed gene.

subnetwork and then discretize the activity score in equally spaced bins. The number of bins is decided by $\lfloor log_2(\text{number of samples}) + 1 \rfloor$. We then calculate mutual information for each of the subnetwork from the discretized mutual information based on class. We pick the protein that has highest mutual information and again expand the subnetwork from recently added protein with $distance = 1$ if the distance from the seed gene to recently added protein is not greater than our intended distance. The process continues till there is no improvement in mutual information or all the proteins are added to the subnetwork or the mutual information is greater than 0.5.

Figure 2.2 shows an example subnetwork with a seed gene protein. The subnetwork mentioned in the figure is a single connected component in the protein protein interaction network. Each protein is mapped to a gene. Activity score that is calculated for a subnetwork is a meta gene. So if we have *k* seed gene protein, we will obtain k subnetworks which means we will obtain k metagene where each gene represents all the samples.

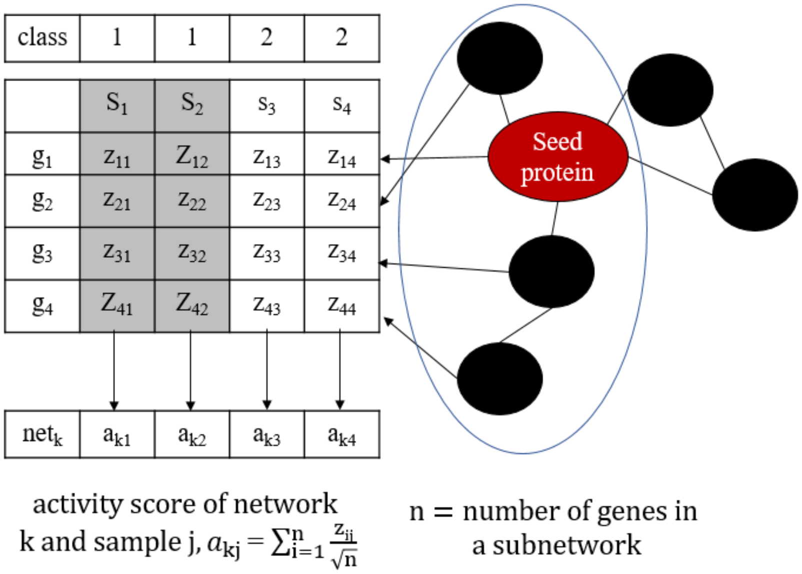

class | 1 | 1 | 2 | 2
$S_1$ $S_2$ $s_3$ $s_4$
$g_1$ $z_{11}$ $Z_{12}$ $z_{13}$ $z_{14}$
$g_2$ $z_{21}$ $z_{22}$ $z_{23}$ $z_{24}$
$g_3$ $z_{31}$ $z_{32}$ $z_{33}$ $z_{34}$
$g_4$ $Z_{41}$ $Z_{42}$ $z_{43}$ $z_{44}$
$net_k$ $a_{k1}$ $a_{k2}$ $a_{k3}$ $a_{k4}$
Seed protein
activity score of network k and sample j, $a_{kj} = \sum_{i=1}^{n} \frac{z_{ij}}{\sqrt{n}}$
n = number of genes in a subnetwork

FIGURE 2.2: Construction of a subnetwork with a seed gene protein.

Figure 2.3 shows the matrix that we will get with $k$ seed gene proteins. Each row is a metagene and each column is the sample of a patient. we used a similar matrix for each subtype to identify network-biomarkers and predict survivability of breast cancer patient.

|  | s1 | s2 | s3 | s4 |
|---|---|---|---|---|
| $subnet_1$ | ▨ | ▨ |  |  |
| $subnet_2$ | ▨ | ▨ |  |  |
| ... | ▨ | ▨ |  |  |
| $subnet_k$ | ▨ | ▨ |  |  |

|  |  |  |  |  |
|---|---|---|---|---|
| class | 1 | 1 | 2 | 2 |

FIGURE 2.3: Matrix of samples and meta genes used for identification of network-biomarkers and patient survival

### Mutual Information

Mutual information for discretized activity score $a'$ and class $c$ is calculated as follows:

$$MI(a', c) = \sum_{x \epsilon a'} \sum_{y \epsilon c} p(x, y) log \frac{p(x, y)}{p(x).p(y)} \tag{2.2}$$

Here $x$ is each element of discretized activity score $a'$ and $y$ is element of each class label $c$. Higher value of mutual information indicates that it can separate the classes more than the lower mutual information.

### Discretization of Activity Score

We discretize the activity score by taking equally spaced bins. Number of bins are decided by $\lfloor log_2(\text{number of samples}) + 1 \rfloor$. The distance of each bin is decided by the following formula:

$$\text{distance of each bin}, \delta = \frac{\text{max value of activity score} - \text{min value of activity score}}{\text{number of bins}} \tag{2.3}$$

## 2.4 Forward Feature Selection

Forward feature selection is a wrapper method for selecting features from the dataset that starts with empty set of features and iteratively add more features. Given a classification algorithm which is random forest in our case, at first step, forward feature selection perform classification $m$ times if there are $m$ features in the dataset and select the best feature. In the next step, it again performs classification $(m - 1)$ times as one feature is previously selected and takes the feature that gives best performance when considered together with previously selected feature or feature-set. The process continues until there is no further improvement after adding additional features. We used 10-fold cross validation and bi-directional search with forward feature selection to select the features that discriminate our two classes. We used Weka [15] machine learning tool and take F1-measure as performance evaluator.

## 2.5 P-Value

P-value indicates the statistical significance of a feature by null hypothesis test. The minimum value is zero and maximum value is 1. Smaller p-value indicates strong evidence against null hypothesis while the larger p-value indicates opposite.

## 2.6 ADASYN

ADASYN [17] is synthetic sample generating technique that generates more samples that are hard to classify by reducing bias and using adaptive learning. Here, adaptive learning means it generates more minority class samples from those minority class samples that are hard to classify. The hard to classify minority class samples are those that have more majority class samples in its k nearest neighbor. The algorithm first calculates the ratio of majority class sample in $K$ nearest neighbor of each minority class sample by using the following formula:

$$r_i = \Delta_i / K \tag{2.4}$$

$\Delta_i$ is the number of majority class samples in $K$ nearest neighbor of $x_i$ where $x_i \epsilon$ minority class sample. Then the algorithm normalize $r_i$ as density distribution as follows:

$$\hat{r_i} = \frac{r_i}{\sum_{i=1}^{m_s} r_i} \tag{2.5}$$

$m_s$ is the minority class samples. Then it calculates the number of samples needs to be generated for each of the minority class samples

$$g_i = \hat{r_i} * G \tag{2.6}$$

$G$ is the total number of samples that needs to be generated. The samples are generated from the following formula:

$$s_i = x_i + (x_{zi} - x_i) * \lambda \tag{2.7}$$

$x_{zi}$ is the minority class samples in K nearest neighbor of $x_i$ and $\lambda$ is a random number between 0 and 1.

We consider patient who decease as positive class and patients who survive as negative class. If the ratio between positive and negative class is greater than or equal to 1:2 or vice-versa, we consider those subsets of the data have class imbalance and apply ADASYN.

## 2.7 Random Forest

Random forest [4] is a tree based classifier that grows multiple classification trees to classify a sample. If there are $n$ samples in the training set, it randomly sample the data from

$n$ samples and grows a number of trees. At each node of the decision trees, it selects $m$ features in our case $m$ genes randomly where $m << M$. $M$ indicates total genes in the dataset. The best split of these $m$ is used to split the node. It then use voting to finally assign class label: decease or survive to the sample. So if most of the trees classify the sample as decease, the algorithm will classify the patient as decease and if most of the trees classify the sample as survive the algorithm will classify the sample as survive. One of the important task of random forest classifier is to optimize the number of features at each split of the decision tree because the classifier's performance may vary because of the number of features at each split. For this purpose, we started with 1 features at each split and, trained and tested with $M-1$ features and chose the model that gives best Geometric-Mean value. We used the implementation of MATLAB's Statistics and Machine Learning Toolbox for random forest.

## 2.8 Logistic Regression

Logistic regression uses logistic or sigmoid function to predict the outcome of a sample. The classifier tries to learn the following objective function:

$$p(y=1|x) = \frac{1}{1+exp(-\theta^t x)} = h_\theta(x) \tag{2.8}$$

$$p(y=0|x) = 1 - p(y=0|x) \tag{2.9}$$

In the equations, $y$ indicates class labels and $x$ are the features. $\theta$ is the parameter we need to find such that $h_\theta(x)$ is large when $x$ belongs to class 1 and small when $x$ belongs to class 2. To find $\theta$ we need to minimize following cost function:

$$J(\theta) = -ylogh_\theta(x) + (1-y)log(1-h_\theta x) \tag{2.10}$$

This cost function indicates how well our hypothesis $h_\theta$ fits our training data.

## 2.9 Our Method

We use chi square feature selection technique to identify the seed genes from gene expression and then using those seed genes we identify the subnetworks from PPI. The identified subnetworks are the features for the next steps. We again use chi square and forward feature selection with random forest to further reduce the number of subnetworks. While using forward feature selection, we use 10-fold cross validation to obtain the features. After identifying the network-biomarkers, we use those network biomarkers to predict the survivability of patients. If the ratio between survived and deceased patients is greater than 2:1, we apply ADASYN to generate synthetic samples for minority class samples and then use random forest classifier to predict the survivability. If the ratio of samples of survived and deceased class is less than 2:1 or vice versa, we directly use random forest classifier. Figure 2.4 shows the steps invovled in our method.

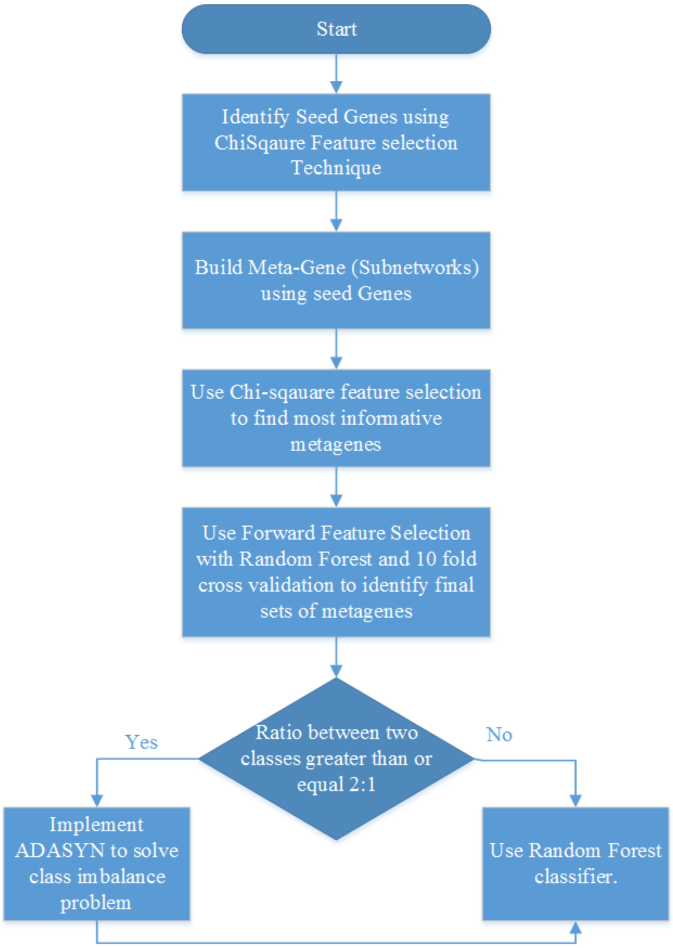

Start
Identify Seed Genes using ChiSqaure Feature selection Technique
Build Meta-Gene (Subnetworks) using seed Genes
Use Chi-sqauare feature selection to find most informative metagenes
Use Forward Feature Selection with Random Forest and 10 fold cross validation to identify final sets of metagenes
Ratio between two classes greater than or equal 2:1
Yes
No
Implement ADASYN to solve class imbalance problem
Use Random Forest classifier.

FIGURE 2.4: Steps of involve in our method

## 2.10 Chuang et al's Method

Seed gene selection in Chuang et al's method is based on common genes of two datasets. As we have only one dataset, we apply Chuang et al's method after seed gene selection procedure. After identifying seed genes, they also identify subnetworks from PPI using seed genes as initial network. Then they apply logistic regression to classify between classes and choose feature based on increasing order of p-value. The classifier starts with one feature and increase the number of features iteratively and finally select all those features that achieve the best performance measure.

In this chapter we describe our method and algorithms used in our method such as chi square, forward feature selection, random forest and metagene identification. We also describe Chuang et al's method and related algorithm to it.

# CHAPTER 3

# *Results*

## 3.1 Identification of Network-biomarkers

First, we identified the seed genes for each subtype and each combination of therapy using Chi square feature selection from gene expression dataset. So the seed genes are actually genes in gene expression data that can discriminate survivability of patients. We consider all the genes as seed genes that have Chi square value greater than zero. Table 3.1 and 3.2 show the number of genes that are identified as seed genes.

TABLE 3.1: Number of seed genes identified for each subtype by using Chi square feature selection.

| **Subtype** | **Features** |
|---|---|
| 1 | 67 |
| 2 | 115 |
| 3 | 142 |
| 4 | 240 |
| 5 | 60 |
| 6 | 86 |
| 7 | 190 |
| 8 | 362 |
| 9 | 69 |
| 10 | 122 |

TABLE 3.2: Number of seed genes identified for each therapy combination by using Chi square feature selection

| **Therapy Combination** | **Features** |
|---|---|
| C | 184 |
| C & H | 417 |
| C & H & R | 131 |
| C & R | 118 |
| H | 127 |
| H & R | 378 |
| R | 136 |
| No Therapy | 537 |

We build the subnetworks using distance 2, 3 and 4 separately for each subtypes and combination of therapy following Chuang et al's [6] approach described in methodology section. These subnetworks are actually meta-genes and we use these metagenes as features. If we have $k$ seed genes we will atleast obtain $k$ subnetworks. We may obtain more than $k$ subnetworks as sometimes a gene can be mapped to more than one protein and in that case we will get more than one subnetworks for one seed gene. Table 3.3 shows the number of subnetworks identified with seed gene.

TABLE 3.3: Number of subnetworks obtained in subtypes.

| **Subtype** | **Seed Genes** |
|---|---|
| 1 | 67 |
| 2 | 115 |
| 3 | 143 |
| 4 | 243 |
| 5 | 62 |
| 6 | 90 |
| 7 | 194 |
| 8 | 369 |
| 9 | 69 |
| 10 | 122 |

TABLE 3.4: Number of subnetworks obtained in combination of therapy.

| **Therapy Combination** | **Seed Genes** |
|---|---|
| **C** | 188 |
| **C & H** | 428 |
| **C & H & R** | 133 |
| **C & R** | 121 |
| **H** | 127 |
| **H & R** | 384 |
| **R** | 140 |
| **No Therapy** | 544 |

With our method, Chi square feature selection technique is applied again to further reduce the number of features but this time we reduce the number of metagenes as we are using subnetworks as our features. After applying Chi square and taking all the metagenes that have Chi square value greater than zero, forward feature selection technique along with random forest classifier is used to identify those metagenes that can discriminate the patients who are going to survive and who are going to be deceased. We took F1-measure as evaluator of our classifier while selecting features with Forward feature selection. Table 3.5 and 3.6 show number of identified network-biomarkers for each subtype and each combination of therapy.

TABLE 3.5: Number of subnetworks for each subtype after applying Chi square and forward feature selection.

| **Subtype** | **Distance 2** | **Distance 3** | **Distance 4** |
|---|---|---|---|
| **1** | 25 | 28 | 30 |
| **2** | 42 | 35 | 34 |
| **3** | 34 | 47 | 41 |
| **4** | 57 | 55 | 60 |
| **5** | 29 | 30 | 31 |
| **6** | 28 | 37 | 38 |
| **7** | 46 | 55 | 54 |
| **8** | 65 | 70 | 73 |
| **9** | 28 | 30 | 29 |
| **10** | 46 | 47 | 55 |

Chuang et al's method also starts with seed gene selection and he used two datasets to find seed genes. As we do not have two datasets, his method is followed after seed gene

TABLE 3.6: Number of subnetworks for each combination of therapy after applying Chi square and forward feature selection.

| Therapy Combination | Distance 2 | Distance 3 | Distance 4 |
|---|---|---|---|
| **C** | 20 | 16 | 17 |
| **C & H** | 6 | 6 | 6 |
| **C & H & R** | 35 | 40 | 39 |
| **C & R** | 38 | 39 | 44 |
| **H** | 47 | 48 | 41 |
| **H & R** | 82 | 88 | 90 |
| **R** | 38 | 41 | 40 |
| **No therapy** | 74 | 84 | 91 |

selection. As subnetwork identification process of our method and his method are similar, we obtain exactly same metagenes. After identifying the metagenes, p-value is calculated for each of the feature and logistic regression classifier is trained and tested with 10-fold cross validation. The features are added according to increasing order of p-values and select those features that gives best performance measure. gmean is used as primary performance measure. Table 3.7 and 3.8 show the number of metagenes selected using Chuang et al approach.

TABLE 3.7: Number of subnetworks for each subtypes after selecting features based on p-values following Chuang's method.

| Subtype | Distance 2 | Distance 3 | Distance 4 |
|---|---|---|---|
| **1** | 3 | 3 | 3 |
| **2** | 6 | 14 | 10 |
| **3** | 4 | 18 | 18 |
| **4** | 81 | 83 | 68 |
| **5** | 8 | 8 | 8 |
| **6** | 4 | 4 | 4 |
| **7** | 8 | 5 | 5 |
| **8** | 28 | 19 | 14 |
| **9** | 53 | 8 | 18 |
| **10** | 19 | 18 | 18 |

## 3.2 Classifier Evalulators

To evaluate our classifier, we used 10 fold cross validation and as performance measure, we used gmean, F1-measure, MCC, AUC and accuracy. As gmean is sensitive to class imbalance and is often preferred over other performance measure if there is class imbalance

TABLE 3.8: Number of subnetworks for each combination of therapy after selecting features based on p-values following Chuang's method

| **Therapy Combination** | **Distance 2** | **Distance 3** | **Distance 4** |
|---|---|---|---|
| **C** | 5 | 22 | 27 |
| **C & H** | 2 | 3 | 3 |
| **C & H & R** | 54 | 2 | 11 |
| **C & R** | 11 | 5 | 9 |
| **H** | 5 | 16 | 4 |
| **H & R** | 56 | 13 | 13 |
| **R** | 11 | 4 | 114 |
| **No therapy** | 6 | 8 | 8 |

within the data, we consider gmean as our primary performance measure. We consider patients who are deceased as positive class and patients who survived as negative class. We also calculated overall performance measure of our classifier.

Accuracy is calculated by dividing all correctly classified breast cancer patient by total samples in the subset. The formula for accuracy is given below:

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \tag{3.1}$$

where TP = True Positive that is the sample is from positive class and it is classified as postive class, FN = False Negative that is the sample is from positive class but it is classified as negative class, FP = False Positive that is the sample is from negative class but it is classified as positive class, TN = True Negative that is the sample is from negative class and it is classified as negative class. F1-measure measures how well the positive class is classified by considering positive predicted value (PPV) and sensitivity. It ranges from 0 to 1 where 1 indicates the perfect classifier. As we consider deceased patients as our positive class, this performance measure will tell us how well we are predicting deceased patients. F1-measure is calculated as follows:

$$\text{F1-measure} = \frac{2 \times PPV \times Sensitivity}{PPV + Sensitivity} \tag{3.2}$$

where

$$\begin{aligned} Sensitivity &= \frac{TP}{TP+FN} \\ PPV &= \frac{TP}{TP+FP} \end{aligned} \tag{3.3}$$

Area Under Curve (AUC) gives us the tradeoff between true positive rate and false positive rate in Receiver Operating Curve space where the values lies between 0 to 1. 1 indicates the perfect classifier. Geometric mean is the square root of sensitivity and specificity. The formula is given below:

$$Gmean = \sqrt{sensitivity \times specificity} \tag{3.4}$$

where sensitivity is calculated using the formula showed in Equation 3.3 and specificity is calculated by:

$$Specificity = \frac{TN}{TN+FP} \tag{3.5}$$

As sensitivity and specificity captures how much samples in positive class and negative class are classified correctly out of all the samples of positive class and negative class respectively, it can be used as primary performance measure when the number of samples in positive and negative class are not equal.

Mathews Correlation Coefficient (MCC) consider TP, TN, FN and FP and is sensitive to class imbalance problem. Unlike other performance measures described above, it ranges from -1 to 1 where 1 means perfect classifier and 0 means average case. The equation is given below:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP) \times (TP+FN) \times (TN+FP) \times (TN+FN)}} \tag{3.6}$$

For calculating the overall performance o classifier based on subtype and therapy combination, we followed Sokolova et. al's approach [27]. We calculated overall TP, TN, FN and FP and then calculated overall performance Measure following above formula.

$$
\begin{aligned}
TP &= \sum_{i=1}^{l} TP_i \\
FN &= \sum_{i=1}^{l} FN_i \\
FP &= \sum_{i=1}^{l} FP_i \\
TN &= \sum_{i=1}^{l} TN_i
\end{aligned}
\tag{3.7}
$$

In the Equation 3.7, $l$ is indicating number of subtypes or number of combination of therapy.

## 3.3 Subtype Specific Survival Prediction

As we consider 10 subtypes, we predicted the survivability based on 10 subtypes and calculated their performance measure based on the technique described above. We implemented our method and Chuang's method in our dataset with distance 2, 3 and 4 from seed gene protein. Table 3.9 to Table 3.9 show the performance measure for each subtype of our method and Chuang et al.'s method. From the tables, we observe that our method has better performance measure in all the cases except subtype five with distance 2 gmean value. Though gmean in subtype five with distance2 is less than Chuang et. al's approach, other two performance measure we consider is larger than their approach. The red colored value in the tables indicate the best performance measure obtained between the two methods.

We also compare how well our method performs overall compared to Chuang et al's method. We observe that for each distance we consider, our method performs better than their method. Table 3.12 shows the comparison of performance measure for each distance. From the table, we observe that with our dataset our method has significantly high performance measure than their method and the difference between the performance measure is almost 10% with all the distances. As we observe that, irrespective of distance, our method performs better than Chuang's method for subtype specific survival prediction, we compare our method between different distances for each subtype to see which distance performs better for which subtype. Table 3.13, 3.14, 3.15, 3.16, 3.17 show gmean, MCC, F1-measure, AUC and accuracy of the classifier respectively. The red colored cells in the table indicate

TABLE 3.9: Comparison of performance measure between Chuang et. al method and our method for subtype specific survival prediction with network distance 2 from seed gene protein.

| | Gmean | | F1-measure | | Accuracy | |
|---|---|---|---|---|---|---|
| **Subtypes** | **Chuang Method** | **Our Method** | **Chuang Method** | **Our Method** | **Chuang Method** | **Our Method** |
| **1** | 0.754 | 0.897 | 0.734 | 0.891 | 75.93% | 89.81% |
| **2** | 0.750 | 1 | 0.816 | 1 | 76.79% | 100% |
| **3** | 0.770 | 0.940 | 0.694 | 0.877 | 84.24% | 92.39% |
| **4** | 0.689 | 0.740 | 0.611 | 0.683 | 71.19% | 79.84% |
| **5** | 0.780 | 0.757 | 0.850 | 0.873 | 80.63% | 82.50% |
| **6** | 0.802 | 0.940 | 0.837 | 0.961 | 80.95% | 95.24% |
| **7** | 0.813 | 0.908 | 0.774 | 0.891 | 82.64% | 91.74% |
| **8** | 0.810 | 0.840 | 0.775 | 0.813 | 81.54% | 85.64% |
| **9** | 0.781 | 0.914 | 0.803 | 0.935 | 78.30% | 92.45% |
| **10** | 0.725 | 0.770 | 0.666 | 0.728 | 75.53% | 81.38% |

TABLE 3.10: Comparison of performance measure between Chuang et. al method and our method for subtype specific survival prediction with network distance 3 from seed gene protein.

| | Gmean | | F1-measure | | Accuracy | |
|---|---|---|---|---|---|---|
| **Subtypes** | **Chuang Method** | **Our Method** | **Chuang Method** | **Our Method** | **Chuang Method** | **Our Method** |
| **1** | 0.731 | 0.908 | 0.718 | 0.903 | 73.15% | 90.74% |
| **2** | 0.777 | 1 | 0.828 | 1 | 78.57% | 100% |
| **3** | 0.777 | 0.927 | 0.686 | 0.859 | 82.61% | 91.30% |
| **4** | 0.653 | 0.733 | 0.568 | 0.675 | 67.49% | 79.42% |
| **5** | 0.767 | 0.790 | 0.833 | 0.896 | 78.75% | 85.63%‘ |
| **6** | 0.802 | 0.947 | 0.837 | 0.960 | 80.95% | 95.24% |
| **7** | 0.807 | 0.860 | 0.766 | 0.829 | 81.82% | 86.78% |
| **8** | 0.801 | 0.857 | 0.764 | 0.834 | 81.03% | 87.18% |
| **9** | 0.794 | 0.914 | 0.826 | 0.935 | 80.19% | 92.45% |
| **10** | 0.748 | 0.774 | 0.695 | 0.734 | 77.13% | 81.91% |

TABLE 3.11: Comparison of performance measure between Chuang et. al method and our method for subtype specific survival prediction with network distance 4 from seed gene protein.

| | **Gmean** | | **F1-measure** | | **Accuracy** | |
|---|---|---|---|---|---|---|
| **Subtype** | **Chuang Method** | **Our Method** | **Chuang Method** | **Our Method** | **Chuang Method** | **Our Method** |
| **1** | 0.731 | 0.907 | 0.718 | 0.902 | 73.15% | 90.74% |
| **2** | 0.763 | 0.986 | 0.811 | 0.985 | 76.79% | 98.21% |
| **3** | 0.770 | 0.940 | 0.694 | 0.897 | 84.24% | 94.02% |
| **4** | 0.678 | 0.725 | 0.598 | 0.662 | 69.14% | 78.19% |
| **5** | 0.767 | 0.797 | 0.833 | 0.895 | 78.75% | 85.63% |
| **6** | 0.802 | 0.960 | 0.837 | 0.973 | 80.95% | 96.83% |
| **7** | 0.807 | 0.884 | 0.766 | 0.860 | 81.82% | 89.26% |
| **8** | 0.788 | 0.825 | 0.750 | 0.794 | 80.51% | 84.10% |
| **9** | 0.795 | 0.906 | 0.803 | 0.926 | 79.25% | 91.51% |
| **10** | 0.748 | 0.782 | 0.695 | 0.742 | 77.13% | 81.91% |

TABLE 3.12: Comparison of overall performance measure between Chuang's method and our method for subtype specific survival prediction.

| **Distancce** | **Performance Measure** | **Chuang Method** | **Our Method** |
|---|---|---|---|
| **2** | **Gmean** | 0.779 | 0.866 |
| | **MCC** | 0.561 | 0.734 |
| | **F1-measure** | 0.751 | 0.849 |
| | **Accuracy** | 78.44% | 86.93% |
| **3** | **Gmean** | 0.770 | 0.867 |
| | **MCC** | 0.542 | 0.735 |
| | **F1-measure** | 0.743 | 0.850 |
| | **Accuracy** | 77.46% | 87.00% |
| **4** | **Gmean** | 0.771 | 0.863 |
| | **MCC** | 0.546 | 0.729 |
| | **F1-Measure** | 0.742 | 0.841 |
| | **Accuracy** | 77.74% | 86.78% |

the best value obtained by the classifier for that specific subtype and specific performance measure. From the tables it is evident that, in most of the subtypes, we obtain best value of different performance measure with distance 3 from the seed gene protein.

TABLE 3.13: Geometric mean for subtype specific survival prediction with our method.

| Subtype | Distance 2 | Distance 3 | Distance 4 |
|---|---|---|---|
| 1 | 0.897 | 0.908 | 0.907 |
| 2 | 1 | 1 | 0.986 |
| 3 | 0.940 | 0.927 | 0.940 |
| 4 | 0.740 | 0.733 | 0.725 |
| 5 | 0.757 | 0.790 | 0.797 |
| 6 | 0.940 | 0.947 | 0.960 |
| 7 | 0.908 | 0.860 | 0.884 |
| 8 | 0.840 | 0.857 | 0.825 |
| 9 | 0.914 | 0.914 | 0.906 |
| 10 | 0.770 | 0.774 | 0.782 |

TABLE 3.14: MCC for subtype specific survival prediction with our method.

| Subtype | Distance 2 | Distance 3 | Distance 4 |
|---|---|---|---|
| 1 | 0.795 | 0.815 | 0.814 |
| 2 | 1 | 1 | 0.962 |
| 3 | 0.832 | 0.807 | 0.857 |
| 4 | 0.550 | 0.540 | 0.512 |
| 5 | 0.615 | 0.692 | 0.689 |
| 6 | 0.904 | 0.901 | 0.935 |
| 7 | 0.825 | 0.721 | 0.773 |
| 8 | 0.699 | 0.732 | 0.667 |
| 9 | 0.850 | 0.850 | 0.830 |
| 10 | 0.602 | 0.615 | 0.613 |

TABLE 3.15: F1-measure for subtype specific survival prediction with our method.

| Subtype | distance 2 | distance 3 | distance 4 |
|---|---|---|---|
| 1 | 0.891 | 0.903 | 0.902 |
| 2 | 1 | 1 | 0.985 |
| 3 | 0.877 | 0.859 | 0.897 |
| 4 | 0.683 | 0.675 | 0.662 |
| 5 | 0.873 | 0.896 | 0.895 |
| 6 | 0.961 | 0.960 | 0.973 |
| 7 | 0.891 | 0.829 | 0.860 |
| 8 | 0.813 | 0.834 | 0.794 |
| 9 | 0.935 | 0.935 | 0.926 |
| 10 | 0.728 | 0.734 | 0.742 |

TABLE 3.16: AUC for subtype specific survival prediction with our method.

| **Subtype** | **Distance 2** | **Distance 3** | **Distance 4** |
|---|---|---|---|
| **1** | 0.918 | 0.965 | 0.968 |
| **2** | 1 | 1 | 0.997 |
| **3** | 0.973 | 0.958 | 0.972 |
| **4** | 0.850 | 0.861 | 0.852 |
| **5** | 0.912 | 0.932 | 0.919 |
| **6** | 0.993 | 0.997 | 0.997 |
| **7** | 0.962 | 0.955 | 0.956 |
| **8** | 0.935 | 0.949 | 0.937 |
| **9** | 0.969 | 0.969 | 0.980 |
| **10** | 0.868 | 0.871 | 0.882 |

TABLE 3.17: Accuracy for subtype specific survival prediction with our method.

| **Subtype** | **Distance 2 (%)** | **Distance 3 (%)** | **Distance 4 (%)** |
|---|---|---|---|
| **1** | 89.81 | 90.74 | 90.74 |
| **2** | 100 | 100 | 98.21 |
| **3** | 92.39 | 91.30 | 94.02 |
| **4** | 79.84 | 79.42 | 78.19 |
| **5** | 82.50 | 85.63 | 85.63 |
| **6** | 95.24 | 95.24 | 96.83 |
| **7** | 91.71 | 86.78 | 89.26 |
| **8** | 85.64 | 87.18 | 84.10 |
| **9** | 92.45 | 92.45 | 91.50 |
| **10** | 81.38 | 81.91 | 81.91 |

We also want to find which distance is the best with our method that gives best performance of classifier. Though we observe that distance 3 has better performance evaluator value in most of the subtypes, without calculating the overall performance measure we cannot say that distance 3 is the best distance. Hence, we calculate overall performance measure for each distance and compare them with each other. Figure 3.1 shows the overall performance measure. Value above each bar indicates the value of the corresponding performance evaluator. The red colored one indicates the best value. We observe that distance 3 performs better than distance 2 and and distance 4 by little margin.

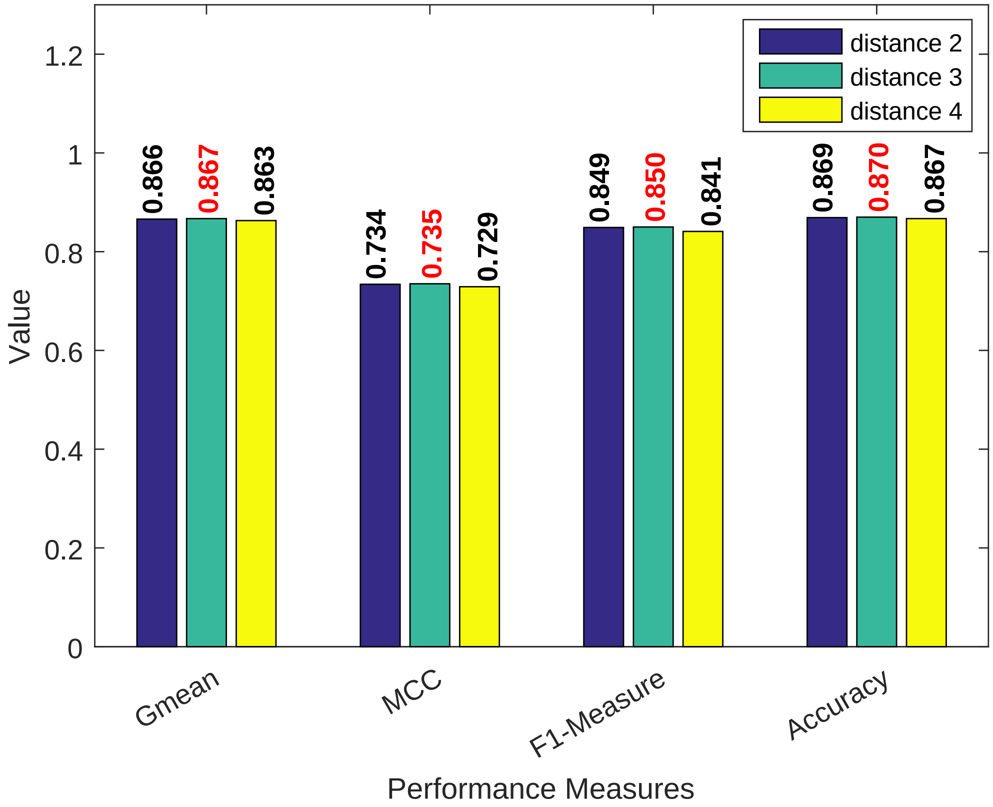


FIGURE 3.1: Subtype specific overall performance measure of the classifiers based on distance with our method.

## 3.4 Therapy Specific Survival Prediction

We have three therapies, that are chemotherapy, hormone therapy and radiotherapy. Thus we got 8 combination of those including the patients did not receive any therapies. Here, we also compare our method with Chuang et al's method. Table 3.18 to Table 3.20 show the performance measure obtained using our approach and Chuang et al's approach.From the tables, we observe that their method gives better value of performance evaluator only for chemotherapy and hormone therapy combination as Chuang et al's method classify all the patients correctly. For rest of the seven therapy combinations, our method outperforms their method as we have better performance measure than theirs. The red colored value in the tables show the best performance measure obtained for that specific therapy combination.

TABLE 3.18: Comparison of performance measure between Chuang et. al method and our method with network distance 2 from seed gene protein for therapy combination.

| | **Gmean** | | **F1-measure** | | **Accuracy** | |
|---|---|---|---|---|---|---|
| **Therapy Combination** | **Chuang Method** | **Our Method** | **Chuang Method** | **Our Method** | **Chuang Method** | **Our Method** |
| **C** | 0.939 | 0.970 | 0.935 | 0.969 | 93.94% | 96.97% |
| **C & H** | 1 | 1 | 1 | 1 | 100% | 100% |
| **C & H & R** | 0.718 | 0.842 | 0.651 | 0.809 | 73.73% | 86.44% |
| **C & R** | 0.788 | 0.866 | 0.740 | 0.842 | 80.15% | 88.55% |
| **H** | 0.713 | 0.804 | 0.666 | 0.772 | 72.58% | 81.61% |
| **H & R** | 0.703 | 0.754 | 0.653 | 0.713 | 72.13% | 77.08% |
| **R** | 0.729 | 0.846 | 0.763 | 0.877 | 73.55% | 85.81% |
| **No therapy** | 0.749 | 0.845 | 0.738 | 0.843 | 75.12% | 84.51% |

TABLE 3.19: Comparison of performance measure between Chuang et. al method and our method with network distance 3 from seed gene protein for therapy combination.

| | **Gmean** | | **F1-measure** | | **Accuracy** | |
|---|---|---|---|---|---|---|
| **Therapy Combination** | **Chuang Method** | **Our Method** | **Chuang Method** | **Our Method** | **Chuang Method** | **Our Method** |
| **C** | 0.939 | 0.970 | 0.941 | 0.969 | 93.94% | 96.97% |
| **C & H** | 1 | 0.925 | 1 | 0.923 | 100% | 94.44% |
| **C & H & R** | 0.637 | 0.897 | 0.552 | 0.873 | 71.19% | 90.68% |
| **C & R** | 0.772 | 0.866 | 0.721 | 0.847 | 79.39% | 89.31% |
| **H** | 0.702 | 0.805 | 0.653 | 0.774 | 71.94% | 81.61% |
| **H & R** | 0.684 | 0.748 | 0.627 | 0.704 | 71.46% | 76.63% |
| **R** | 0.716 | 0.846 | 0.751 | 0.877 | 72.26% | 85.81% |
| **No therapy** | 0.736 | 0.859 | 0.728 | 0.855 | 73.71% | 85.92% |

We calculate the overall performance measure to find which method performs well for each distance. Table 3.21 shows the overall comparison for different distance between our method and their method. We observe that irrespective of any distance, our method performs better than their method for therapy combination. In most of the performance

TABLE 3.20: Comparison of performance measure between Chuang et. al method and our method with network distance 4 from seed gene protein for therapy combination.

| | **Gmean** | | **F1-measure** | | **Accuracy** | |
|---|---|---|---|---|---|---|
| **Therapy Combination** | **Chuang Method** | **Our Method** | **Chuang Method** | **Our Method** | **Chuang Method** | **Our Method** |
| **C** | 0.909 | 0.970 | 0.909 | 0.969 | 90.91% | 96.97% |
| **C & H** | 1 | 0.925 | 1 | 0.923 | 100% | 94.44% |
| **C & H & R** | 0.661 | 0.867 | 0.582 | 0.837 | 72.03% | 88.14% |
| **C & R** | 0.750 | 0.888 | 0.693 | 0.872 | 77.10% | 90.84% |
| **H** | 0.686 | 0.810 | 0.629 | 0.780 | 71.94% | 82.58% |
| **H & R** | 0.684 | 0.760 | 0.627 | 0.721 | 71.46% | 78.65% |
| **R** | 0.727 | 0.836 | 0.750 | 0.874 | 72.90% | 85.16% |
| **No therapy** | 0.736 | 0.853 | 0.728 | 0.848 | 73.71% | 85.45% |

evaluator, our method gives almost 10% better performance evaluator value than theirs. From our observation, we can say that with our dataset, our method outperforms Chuang et al's method for predicting survivability based on therapy combination.

TABLE 3.21: Comparison of overall performance measure between Chuang's method and our method for therapy specific survivability prediction.

| **Distance** | **Performance Measure** | **Chuang Method** | **Our Method** |
|---|---|---|---|
| **2** | **Gmean** | 0.738 | 0.821 |
| | **MCC** | 0.481 | 0.647 |
| | **F1-measure** | 0.704 | 0.799 |
| | **Accuracy** | 74.56% | 82.70% |
| **3** | **Gmean** | 0.723 | 0.825 |
| | **MCC** | 0.459 | 0.656 |
| | **F1-measure** | 0.684 | 0.803 |
| | **Accuracy** | 73.58% | 83.10% |
| **4** | **Gmean** | 0.718 | 0.829 |
| | **MCC** | 0.455 | 0.699 |
| | **F1-measure** | 0.678 | 0.807 |
| | **Accuracy** | 73.44% | 83.80% |

We built our network with three different distances from seed genes. Table 3.22 to Table 3.26 show the comparison of performance measure we estimated using 10-fold cross validation for different distances. From these tables, we can identify which distance performs better for a specific therapy combination. We also observe that distance 3 and distance 4 outperforms distance 2 in most of the therapy combination. The red colored values indicates the highest performance evaluated for the specific therapy combination.

We want to identify which distance gives better performance overall for predicting sur-

TABLE 3.22: Gmean for each combination of therapy with our method.

| **Therapy** | **Distance 2** | **Distance 3** | **Distance 4** |
|---|---|---|---|
| **C** | 0.970 | 0.970 | 0.970 |
| **C & H** | 1 | 0.925 | 0.925 |
| **C & H & R** | 0.842 | 0.897 | 0.867 |
| **C & R** | 0.866 | 0.866 | 0.888 |
| **H** | 0.804 | 0.805 | 0.810 |
| **H & R** | 0.754 | 0.746 | 0.760 |
| **R** | 0.846 | 0.846 | 0.836 |
| **No Therapy** | 0.845 | 0.859 | 0.853 |

TABLE 3.23: MCC for each combination of therapy with our method.

| **Therapy** | **Distance 2** | **Distance 3** | **Distance 4** |
|---|---|---|---|
| **C** | 0.970 | 0.970 | 0.970 |
| **C & H** | 1 | 0.925 | 0.925 |
| **C & H & R** | 0.842 | 0.897 | 0.867 |
| **C & R** | 0.866 | 0.866 | 0.888 |
| **H** | 0.804 | 0.805 | 0.810 |
| **H & R** | 0.754 | 0.746 | 0.760 |
| **R** | 0.846 | 0.846 | 0.836 |
| **No Therapy** | 0.845 | 0.859 | 0.853 |

TABLE 3.24: F1-measure for each combination of therapy with our method.

| **Therapy** | **Distance 2** | **Distance 3** | **Distance 4** |
|---|---|---|---|
| **C** | 0.969 | 0.969 | 0.969 |
| **C & H** | 1 | 0.923 | 0.923 |
| **C & H & R** | 0.809 | 0.873 | 0.837 |
| **C & R** | 0.842 | 0.847 | 0.872 |
| **H** | 0.772 | 0.774 | 0.780 |
| **H & R** | 0.713 | 0.704 | 0.721 |
| **R** | 0.877 | 0.877 | 0.874 |
| **No Therapy** | 0.843 | 0.855 | 0.848 |

TABLE 3.25: AUC for each combination of therapy with our method.

| Therapy | Distance 2 | Distance 3 | Distance 4 |
|---|---|---|---|
| **C** | 1 | 1 | 1 |
| **C & H** | 1 | 1 | 1 |
| **C & H & R** | 0.918 | 0.935 | 0.935 |
| **C & R** | 0.957 | 0.953 | 0.946 |
| **H** | 0.874 | 0.860 | 0.893 |
| **H & R** | 0.825 | 0.820 | 0.820 |
| **R** | 0.925 | 0.925 | 0.939 |
| **No Therapy** | 0.902 | 0.902 | 0.896 |

TABLE 3.26: Accuracy for each combination of therapy with our method.

| Therapy | Distance 2 (%) | Distance 3 (%) | Distance 4 (%) |
|---|---|---|---|
| **C** | 96.97 | 96.97 | 96.97 |
| **C & H** | 1 | 94.44 | 94.44 |
| **C & H & R** | 86.44 | 90.68 | 88.14 |
| **C & R** | 88.55 | 89.31 | 90.84 |
| **H** | 81.61 | 81.61 | 82.58 |
| **H & R** | 77.08 | 76.63 | 78.65 |
| **R** | 85.81 | 85.81 | 85.16 |
| **No Therapy** | 84.51 | 85.92 | 85.45 |

vivability of breast cancer patients based on therapy. Hence, we also calculated overall performance measure for each distance. Figure 3.2 shows the overall performance measure . From the figure, it is evident that distance 4 gives the best performance while distance 2 is the lowest one among the three. So distance 4 is the best distance among the three distance for predicting therapy specific survivability of breast cancer patients.

## 3.5 Comparison between therapy and subtype specific method

If we compare our result between therapy specific survival prediction and subtype specific survival prediction, we observe that subtype specific can better predicts the survivability of breast cancer patients. Table 3.27 shows the comparison between them. We compare between distance 3 subtype specific survivial and distance 4 therapy specific survival as these two are the best in their own category with our method. While subtype specific method has gmean: 0.867, MCC: 0.735, F1-measure: 0.850 and accuracy: 87.00%, therapy specific has gmean: 0.829, MCC: 0.669, F1-measure: 0.807 and accuracy: 83.77%.

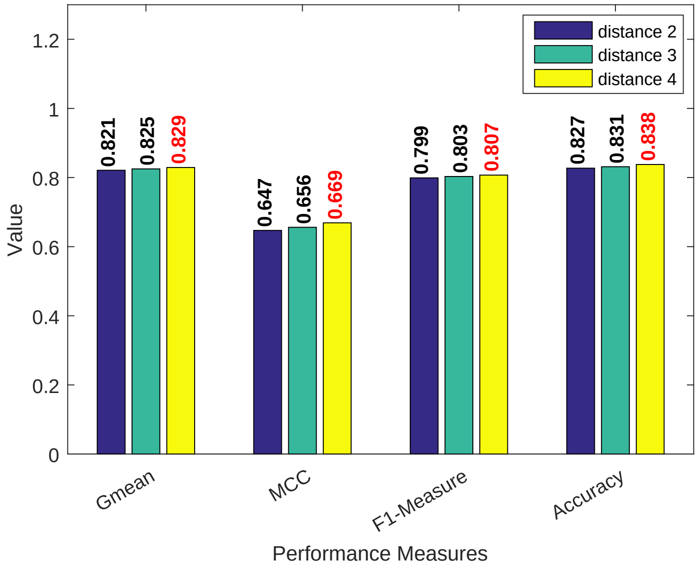


FIGURE 3.2: Therapy specific overall performance measure of the classifiers based on distance with our method.

TABLE 3.27: Comparison of performance measure between subtype specific and therapy specific survival prediction with our method.

| **Performance Measure** | **Subtype Specific** | **Therapy Specific** |
|---|---|---|
| **Gmean** | 0.867 | 0.829 |
| **MCC** | 0.735 | 0.669 |
| **F1-measure** | 0.850 | 0.807 |
| **Accuracy** | 87.00% | 83.77% |

## 3.6 Identified Network biomarkers

We identified medium sized network-biomarkers where the maximum number of protein in a subnetwork for subtype specific survival with distance 3 from seed gene protein is 77 and for therapy specific survival with distance 4 from seed gene protein is 198. Figure A.31 shows the network-biomarkers with distance 2 from seed gene protein. The self loop indicates the seed gene protein. Protein name is mentioned beside each nodes of the network. Figure 3.4 shows the mapped gene to its corresponding proteins. The maximum size of these subnetworks is 31. Rest of the subnetworks are added in the appendix. As some of the network-biomarkers are not clear from the figure, we added the adjacency matrix of the network-biomarkers in supplementary [1].

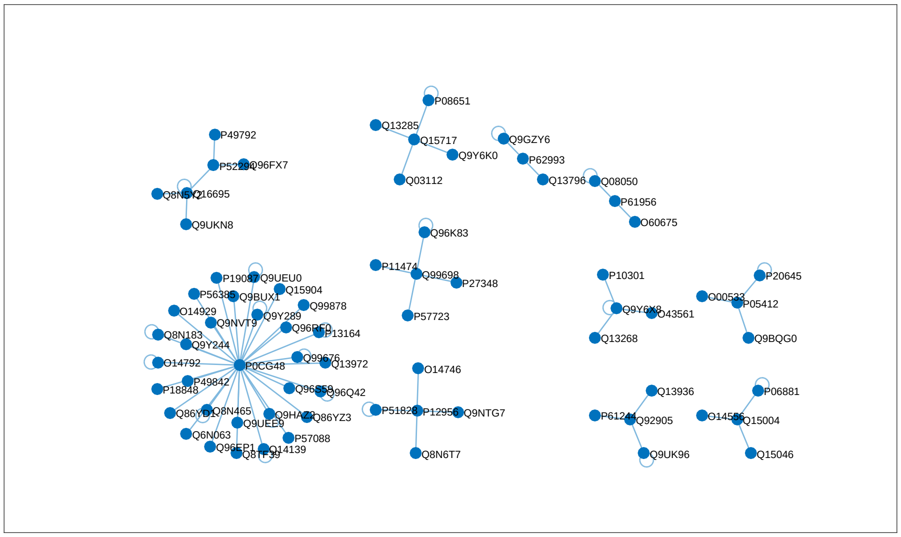


FIGURE 3.3: Identified network-biomarkers for chemotherapy received patients with distance 2 from seed gene proteins where nodes are proteins.

In this chapter, we discuss how we identify the network-biomarkers, compare between our method and Chuang et. al's method and also compare our method among different distances of network from seed gene protein. We also compare between subtype specific survival prediction and therapy specific survival prediction.

[1] http://www.sheikhjubair.com/wp-content/uploads/2017/04/adjacencymatrices.zip

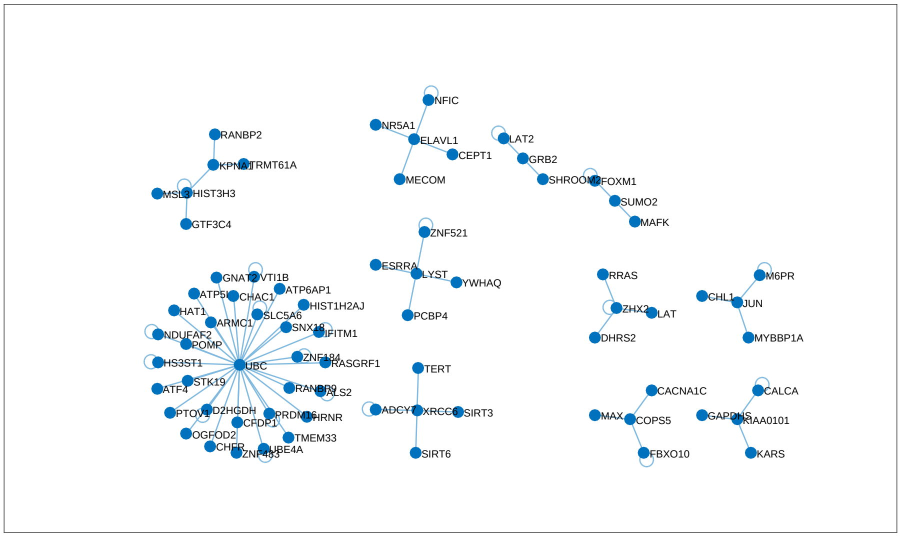


FIGURE 3.4: Identified network-biomarkers for chemotherapy received patients with distance 2 from seed gene proteins where nodes are mapped genes to proteins.

# CHAPTER 4

# *Discussion*

As gene biomarkers are not stable and network-biomarkers based on functionally related genes give stability to the biomarkers, we used a metagene based method where the genes in a metagene are connected through protein protein interaction network. To build the network, we initially start with seed genes and we take all genes as seed genes that have chi square value greater than zero. This gives us all the genes that shows discriminative behavior for predicting survivability of breast cancer patient. With these seed genes, we then build metagenes using protein protein interaction network. Number of genes in a metagene depends on the threshold of mutual information. We continue to add more genes to the metagene until mutual information value goes above 0.5. If we decrease the threshold to some other value, it will build small metagenes than ours. Increasing the threshold may work as just opposite. We then use chi square feature selection technique to reduce the number of metagenes for each subtype and each therapy combination and then use forward feature selection with random forest classifier to obtain the network-biomarkers that can predict whether a patient is going to survive or not. We use forward feature selection after chi square feature selection because forward feature selection is computationally expensive. If the number of features are high, it will take a lot of time to select features that discriminate the classes. We also solve the class imbalance within the data if the ratio of samples between two classes are greater than 1 : 2. If number of samples are not equal, classifiers tend to become biased towards the class that has more samples. ADASYN helps us to solve this problem as we applied ADASYN to generate synthetic samples. For predicting survivability, we used a tree based classifier random forest which also works well if the samples of two classes are not equal.

After obtaining metagenes we used chi square and forward feature selection for selecting discriminative metagenes for our classes and use random forest as classifier while Chuang et. al used p-value for selecting features and logistic regression as classifier. We also address the class imbalance within the subset of the data. Our method finds better discriminative features as we used combination of filter and wrapper method. Logistic regression uses sigmoid function while random forest is based on multiple decision trees. In this dataset, the decision trees with the features we selected discriminate the classes better than sigmoid function based logistic regression where p-value is used to rank the features.

We identified a number of metagenes for subtype and therapy specific survivability prediction. We observe that subtype specific survivability prediction performs better than therapy specific survivability prediction. This outcome is expected as each subtype of breast cancer has different molecular mechanisms and these molecular mechanisms are not considered when we consider therapy combination only.

In our dataset, we have very few samples for both classes in chemotherapy and, chemotherapy and hormonetherapy combination. In chemotherapy, we have 17 patients who survived and 16 patients who are deceased. In chemotherapy and hormonetherapy combination, we have 11 patients who survived and 7 patients who are deceased. As we have very few samples in these two therapy combination, our classifier may suffer from under-fitting problem. As our identified network-biomarkers are of medium size, number of genes that are needed to discriminate survivability of patients are very high compared to gene expression based study. This problem can be solved by optimizing mutual information value for metagene identification. While predicting survivability based on subtypes, we did not consider therapy combination as we have very few samples for different therapy combination in a subtype. If we could consider therapy combination within a subtype, that may gives us better result for finding therapy for a patient.

# CHAPTER 5

# *Conclusion*

In this thesis, we devised a network based machine learning approach that can identify network-biomarkers for predicting survivability of breast cancer patient a) based on subtypes and b) based on therapy. Our method can also predict the survivability of breast cancer patient based on a) subtypes and b) combination of therapy. our method is based on a) integrating gene expression data with protein protein interaction network that allows us to capture the knowledge required for predicting survivability b) devising an algorithm which find the subnetworks that can best discriminate the classes and c) apply machine learning method using the identified network-biomarkers to predict survivability. We use distance 2, 3 and 4 from seed gene protein to identify network biomarkers and compare our method with Chuang et al's method and observe that our method obtain better performance measure by almost 10% in all the cases. We also compare our method between different distances we take for seed gene. We achieved 0.867, 0.735, 0.850 and 87.00% Geometric Mean, Mathews Correlation Coefficient, F1-measure and accuracy for subtype specific with distance 3 and 0.829, 0.669, 0.807, 83.77% Geometric Mean, Mathews Correlation Coefficient, F1-measure and accuracy for distance 4 for therapy specific prediction of breast cancer survivability. As subtypes of breast cancer represents more molecular mechanism, we also observe that subtype specific prediction obtains better result than therapy specific prediction for predicting survivability of patients.

## 5.1 Future Work

Our method can be further improved if we can consider therapy combination for each subtype and predict the survivability based on therapy for each subtype. For this purpose, we need a larger dataset that has enough samples for each combination of therapy. Another way to improve our method is to add more information to our metagenes by considering gene's mutation information or by adding secondary network. Gene's mutation information can be obtained from CNA. As a secondary network, we can consider some predefined biomolecular networks such as co-expression network, cellular pathway maps and regulatory network motifs.

# REFERENCES

[1] (2017). cBioPortal. `http://www.cbioportal.org/`. Online; accessed 10 April 2017.

[2] Akker, E. v. d., Verbruggen, B., Heijmans, B., Beekman, M., Kok, J., Slagboom, E., and Reinders, M. (2011). Integrating protein-protein interaction networks with gene-gene co-expression networks improves gene signatures for classifying breast cancer metastasis. *Journal of Integrative Bioinformatics (JIB)*, 8(2):222–238.

[3] Allahyar, A. and de Ridder, J. (2015). Feral: network-based classifier with application to breast cancer outcome prediction. *Bioinformatics*, 31(12):i311–i319.

[4] Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

[5] Carey, L. A., Dees, E. C., Sawyer, L., Gatti, L., Moore, D. T., Collichio, F., Ollila, D. W., Sartor, C. I., Graham, M. L., and Perou, C. M. (2007). The triple negative paradox: primary tumor chemosensitivity of breast cancer subtypes. *Clinical Cancer Research*, 13(8):2329–2334.

[6] Chuang, H.-Y., Lee, E., Liu, Y.-T., Lee, D., and Ideker, T. (2007). Network-based classification of breast cancer metastasis. *Molecular Systems Biology*, 3(1):140.

[7] Curtis, C., Shah, S. P., Chin, S.-F., Turashvili, G., Rueda, O. M., Dunning, M. J., Speed, D., Lynch, A. G., Samarajiwa, S., Yuan, Y., et al. (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486(7403):346–352.

[8] Dao, P., Wang, K., Collins, C., Ester, M., Lapuk, A., and Sahinalp, S. C. (2011). Optimally discriminative subnetwork markers predict response to chemotherapy. *Bioinformatics*, 27(13):i205–i213.

[9] 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.

[10] Ein-Dor, L., Kela, I., Getz, G., Givol, D., and Domany, E. (2005). Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics*, 21(2):171–178.

[11] Garcia, M., Millat-Carus, R., Bertucci, F., Finetti, P., Guille, A., Adelaıde, J., Bekhouche, I., Sabatier, R., Chaffanet, M., Birnbaum, D., et al. (2014). 12 cnv-interactome-transcriptome integration to detect driver genes in cancerology. *Microarray Image and Data Analysis: Theory and Practice*, page 313.

[12] Glass, K., Quackenbush, J., Spentzos, D., Haibe-Kains, B., and Yuan, G.-C. (2015). A network model for angiogenesis in ovarian cancer. *BMC Bioinformatics*, 16(1):115.

[13] Greenman, C., Stephens, P., Smith, R., Dalgliesh, G. L., Hunter, C., Bignell, G., Davies, H., Teague, J., Butler, A., Stevens, C., et al. (2007). Patterns of somatic mutation in human cancer genomes. *Nature*, 446(7132):153–158.

[14] Hahn, M. E. and MacLean, M. S. (1955). Prognosis and prediction.

[15] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The weka data mining software: An update. *SIGKDD Explor. Newsl.*, 11(1):10–18.

[16] Hanahan, D. and Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *Cell*, 144(5):646–674.

[17] He, H., Bai, Y., Garcia, E. A., and Li, S. (2008). Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *Neural Networks, 2008. IJCNN 2008.(IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on*, pages 1322–1328. IEEE.

[18] Hua, J., Tembe, W. D., and Dougherty, E. R. (2009). Performance of feature-selection methods in the classification of high-dimension data. *Pattern Recognition*, 42(3):409–424.

[19] Lamond, N. W., Skedgel, C., and Younis, T. (2013). Is the 21-gene recurrence score a cost-effective assay in endocrine-sensitive node-negative breast cancer? *Expert Review of Pharmacoeconomics & Outcomes Research*, 13(2):243–250.

[20] Lee, E., Chuang, H.-Y., Kim, J.-W., Ideker, T., and Lee, D. (2008). Inferring pathway activity toward precise disease classification. *PLoS Comput Biology*, 4(11):e1000217.

[21] Merikangas, K. R., Low, N. C., and Hardy, J. (2006). Commentary: Understanding sources of complexity in chronic diseases—the importance of integration of genetics and epidemiology. *International Journal of Epidemiology*, 35(3):590.

[22] Milioli, H. H., Vimieiro, R., Riveros, C., Tishchenko, I., Berretta, R., and Moscato, P. (2015). The discovery of novel biomarkers improves breast cancer intrinsic subtype prediction and reconciles the labels in the metabric data set. *PloS One*, 10(7):e0129711.

[23] Network, C. G. A. et al. (2012). Comprehensive molecular portraits of human breast tumors. *Nature*, 490(7418):61.

[24] Sawyers, C. L. (2008). The cancer biomarker problem. *Nature*, 452(7187):548–552.

[25] Schneider, J. E., Sidhu, M. K., Doucet, C., Kiss, N., Ohsfeldt, R. L., and Chalfin, D. (2012). Economics of cancer biomarkers. *Personalized Medicine*, 9(8):829–837.

[26] Sheikh Jubair (2017). Adjacency Matrices. `http://www.sheikhjubair.com/wp-content/uploads/2017/04/adjacencymatrices.zip`. Online; accessed 10 April 2017.

[27] Sokolova, M. and Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4):427–437.

[28] Soneson, C., Gerster, S., and Delorenzi, M. (2014). Batch effect confounding leads to strong bias in performance estimates obtained by cross-validation. *PloS One*, 9(6):e100335.

[29] Symmans, W. F., Liu, J., Knowles, D. M., and Inghirami, G. (1995). Breast cancer heterogeneity: evaluation of clonality in primary and metastatic lesions. *Human Pathology*, 26(2):210–216.

[30] Umetani, N., Giuliano, A. E., Hiramatsu, S. H., Amersi, F., Nakagawa, T., Martino, S., and Hoon, D. S. (2006). Prediction of breast tumor progression by integrity of free circulating dna in serum. *Journal of Clinical Oncology*, 24(26):4270–4276.

[31] 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.

[32] Verma, M. and Manne, U. (2006). Genetic and epigenetic biomarkers in cancer diagnosis and identifying high risk populations. *Critical Reviews in Oncology/Hematology*, 60(1):9–18.

[33] Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz, L. A., and Kinzler, K. W. (2013). Cancer genome landscapes. *Science*, 339(6127):1546–1558.

[34] Wang, Y., Klijn, J. G., Zhang, Y., Sieuwerts, A. M., Look, M. P., Yang, F., Talantov, D., Timmermans, M., Meijer-van Gelder, M. E., Yu, J., et al. (2005). Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *The Lancet*, 365(9460):671–679.

[35] Wood, L. D., Parsons, D. W., Jones, S., Lin, J., Sjöblom, T., Leary, R. J., Shen, D., Boca, S. M., Barber, T., Ptak, J., et al. (2007). The genomic landscapes of human breast and colorectal cancers. *Science*, 318(5853):1108–1113.

[36] Wu, D., Rice, C. M., and Wang, X. (2012). Cancer bioinformatics: A new approach to systems clinical medicine. *BMC Bioinformatics*, 13(1):71.

[37] Zhang, K. X. and Ouellette, B. F. (2011). Caerus: predicting cancer outcomes using relationship between protein structural information, protein networks, gene expression data, and mutation data. *PLoS Comput Biology*, 7(3):e1001114.

# APPENDIX A

# *Figures*

## A.1 Subnetworks for Subtype Specific Survival Prediction

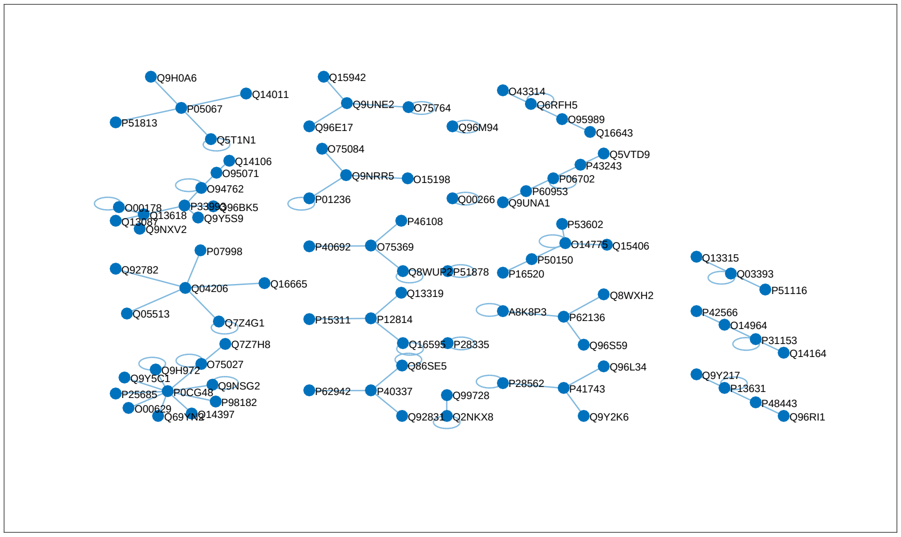


FIGURE A.1: Identified network-biomarkers for subtype 1 distance 2 from seed gene proteins. The maximum size of a subnetwork is 11.

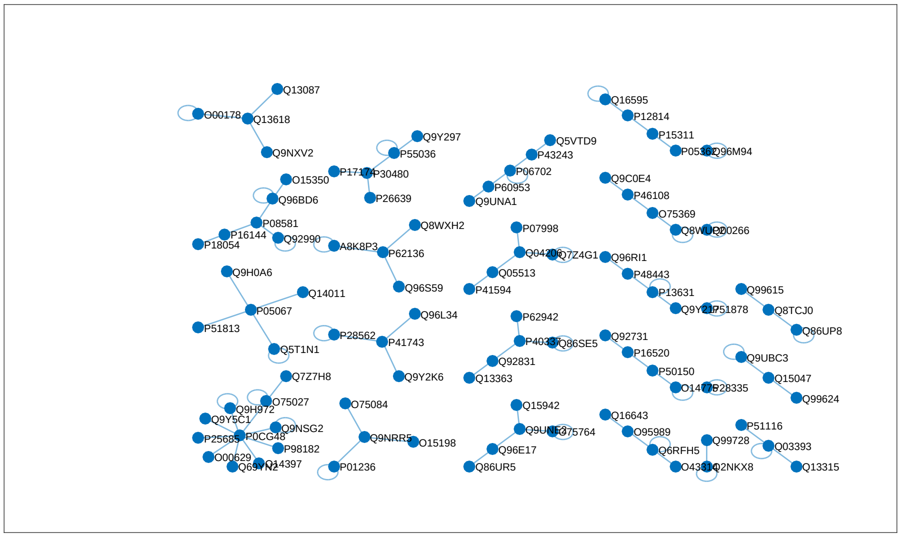


FIGURE A.2: Identified network-biomarkers for subtype 1 distance 3 from seed gene proteins. The maximum size of a subnetwork is 11.

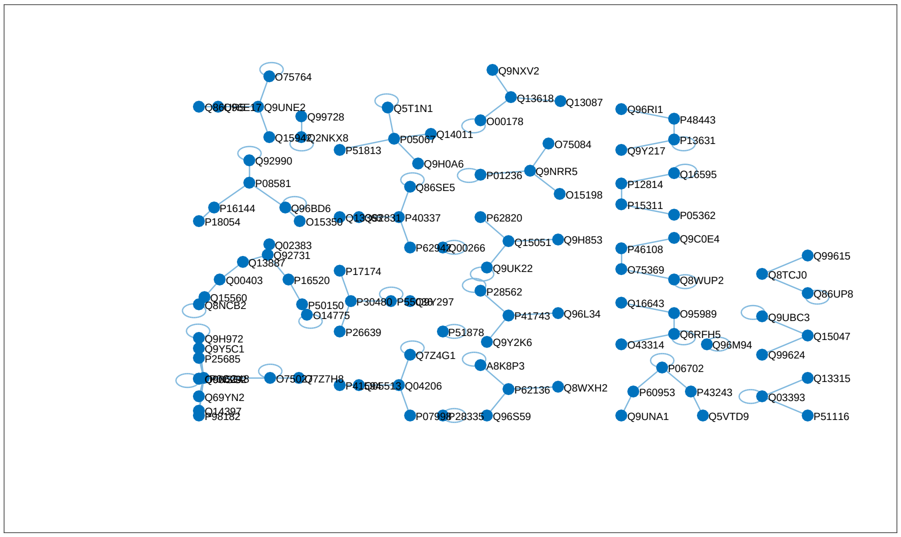


FIGURE A.3: Identified network-biomarkers for subtype 1 distance 4 from seed gene proteins. The maximum size of a subnetwork is 11.

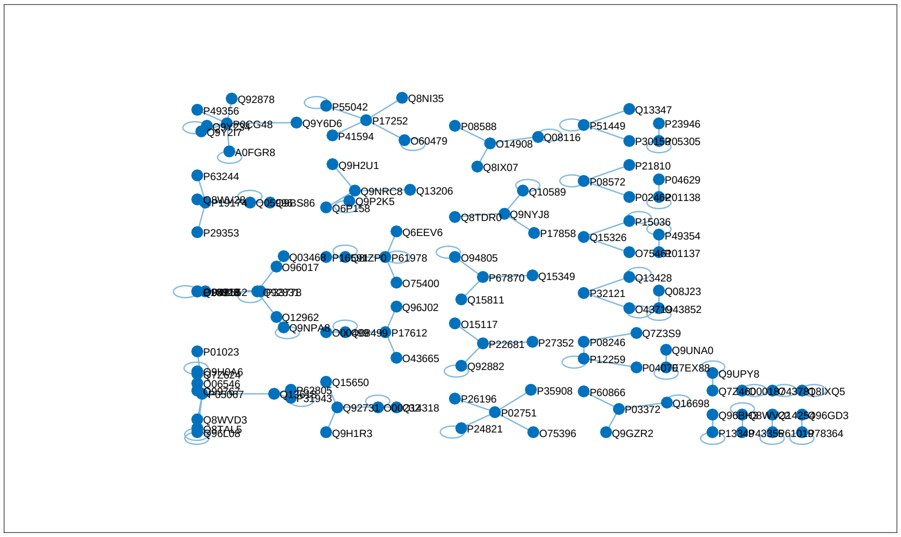


FIGURE A.4: Identified network-biomarkers for subtype 2 distance 2 from seed gene proteins. The maximum size of a subnetwork is 12.

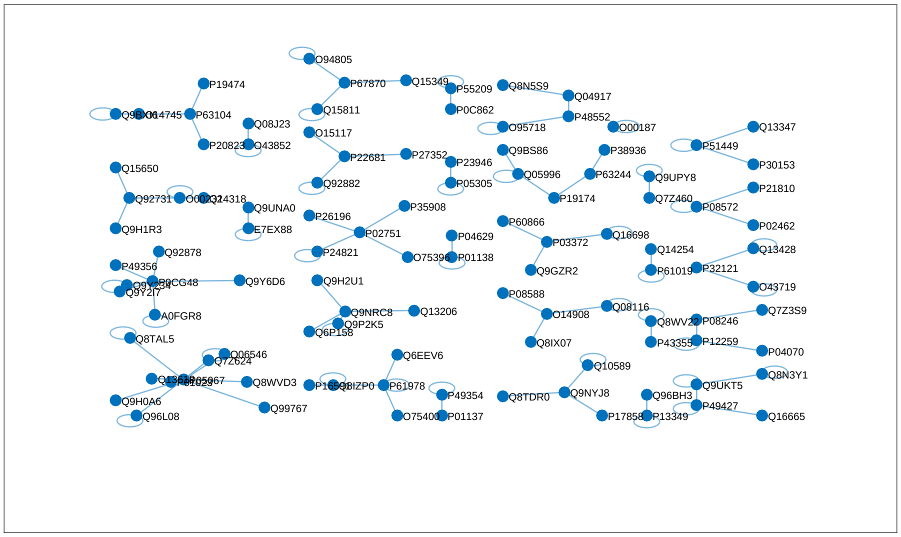


FIGURE A.5: Identified network-biomarkers for subtype 2 distance 3 from seed gene proteins. The maximum size of a subnetwork is 10.

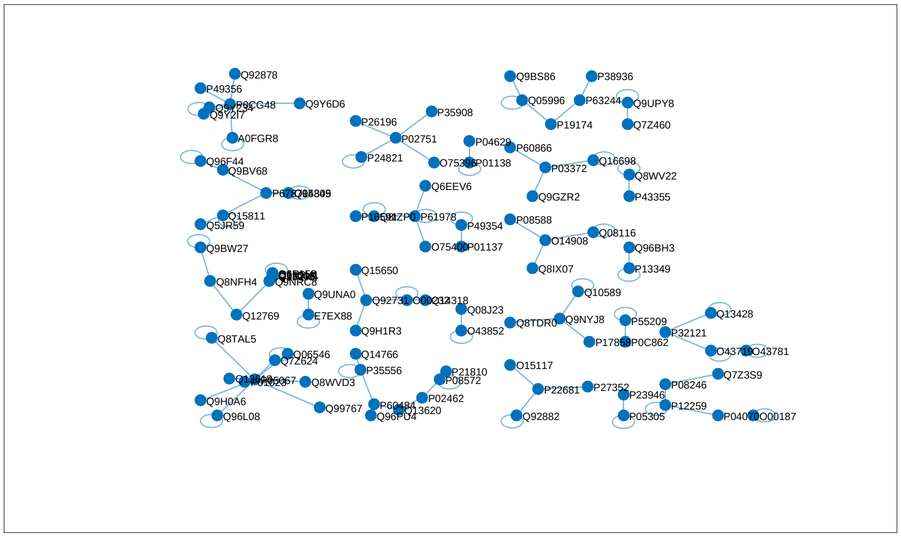


FIGURE A.6: Identified network-biomarkers for subtype 2 distance 4 from seed gene proteins. The maximum size of a subnetwork is 10.

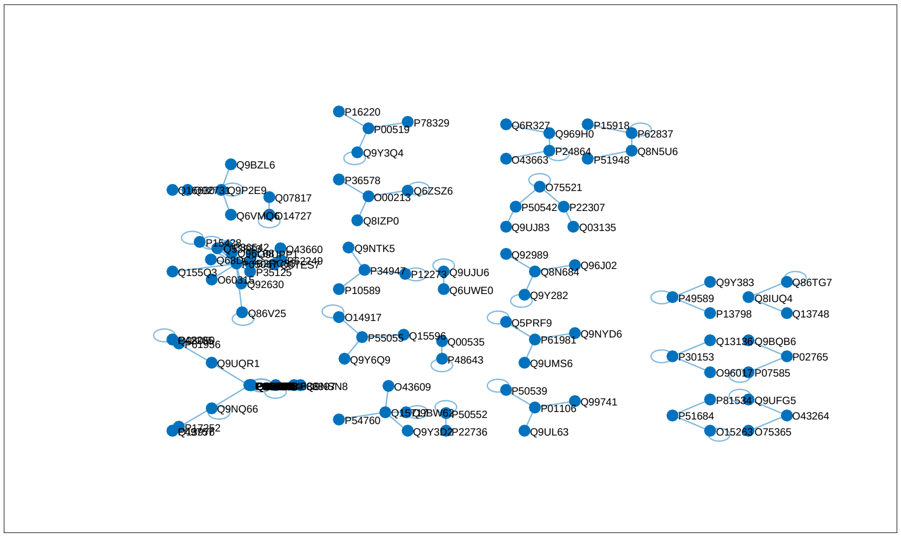


FIGURE A.7: Identified network-biomarkers for subtype 3 distance 2 from seed gene proteins. The maximum size of a subnetwork is 30.

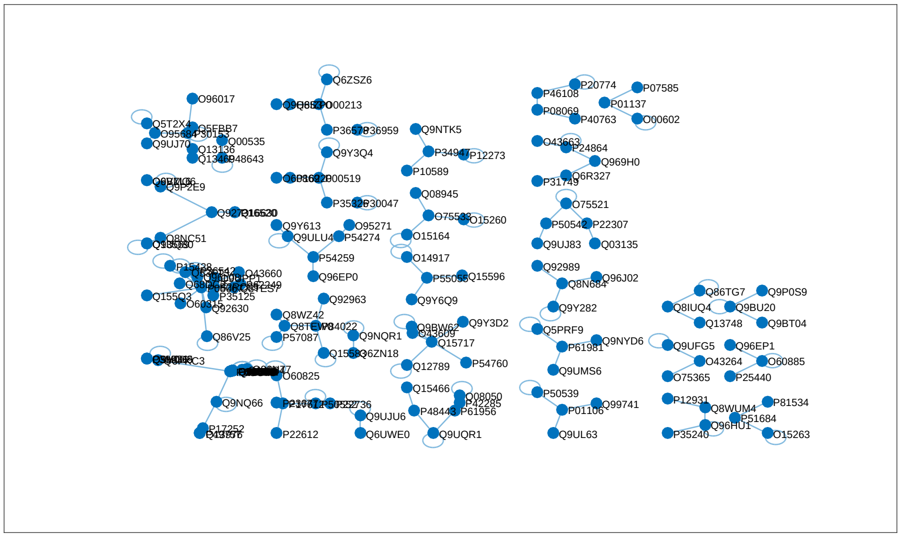


FIGURE A.8: Identified network-biomarkers for subtype 3 distance 3 from seed gene proteins. The maximum size of a subnetwork is 28.

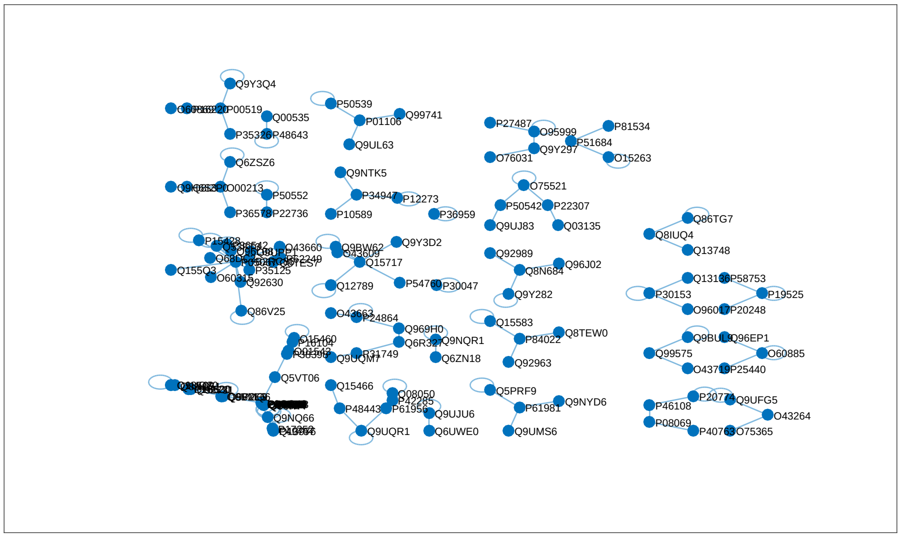


FIGURE A.9: Identified network-biomarkers for subtype 3 distance 4 from seed gene proteins. The maximum size of a subnetwork is 37.

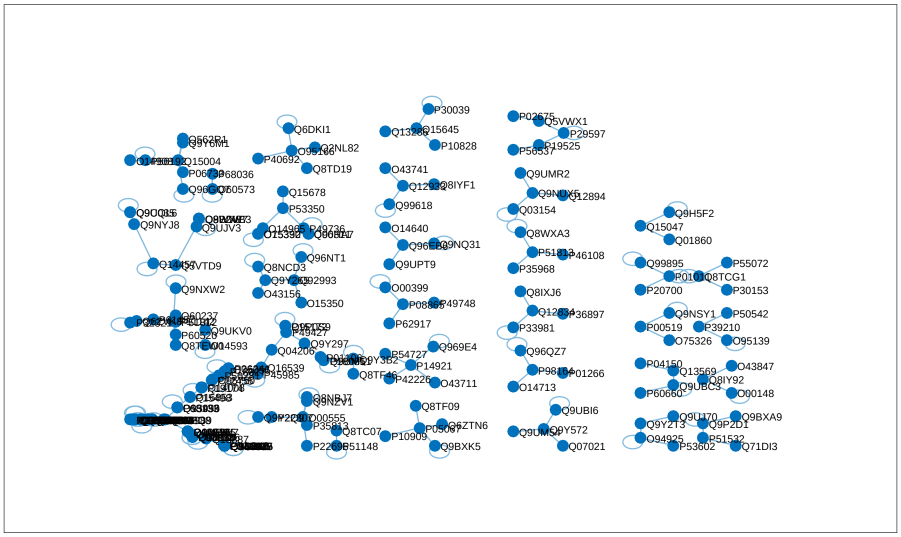


FIGURE A.10: Identified network-biomarkers for subtype 4 distance 2 from seed gene proteins. The maximum size of a subnetwork is 53.

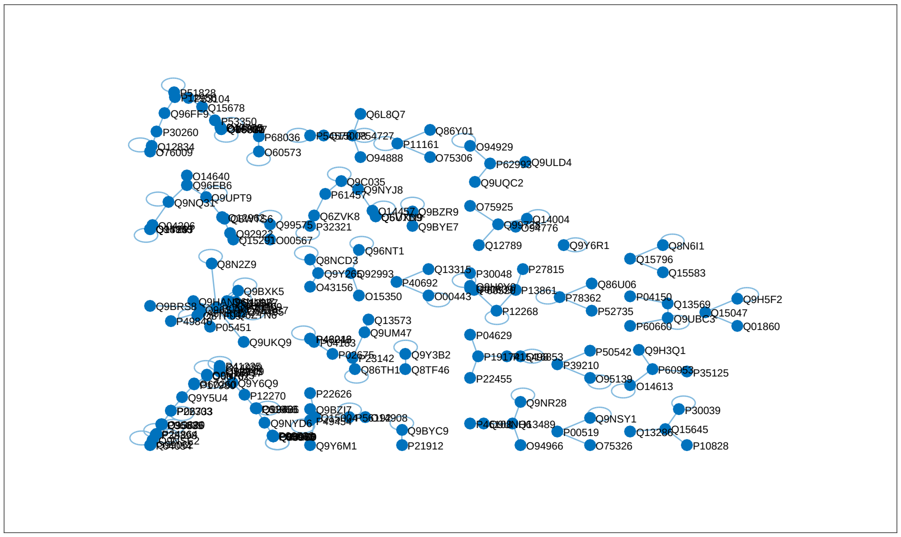


FIGURE A.11: Identified network-biomarkers for subtype 4 distance 3 from seed gene proteins. The maximum size of a subnetwork is 37.

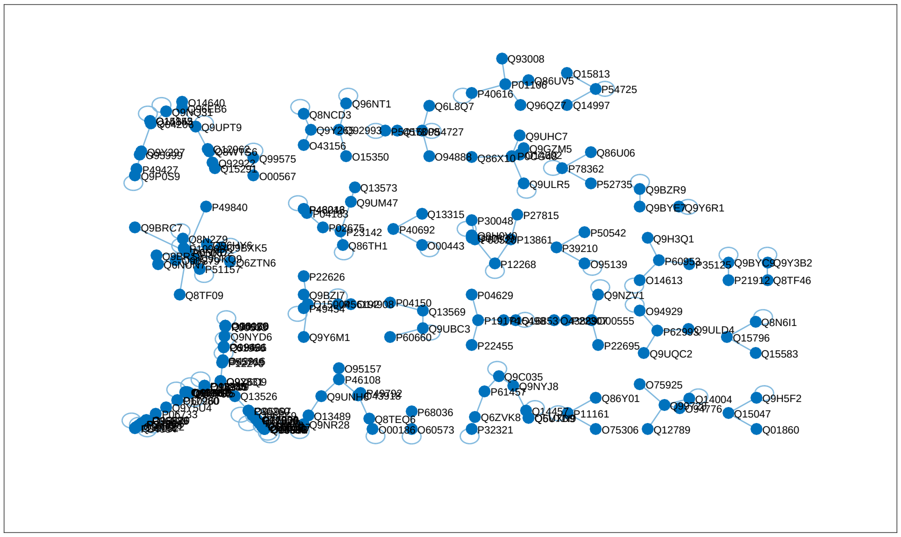


FIGURE A.12: Identified network-biomarkers for subtype 4 distance 4 from seed gene proteins. The maximum size of a subnetwork is 56.

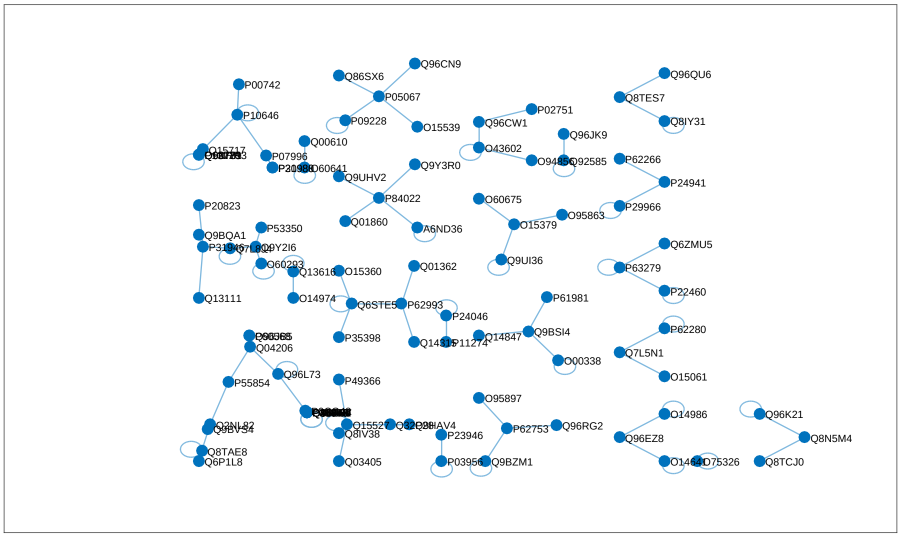


FIGURE A.13: Identified network-biomarkers for subtype 5 distance 2 from seed gene proteins. The maximum size of a subnetwork is 22.

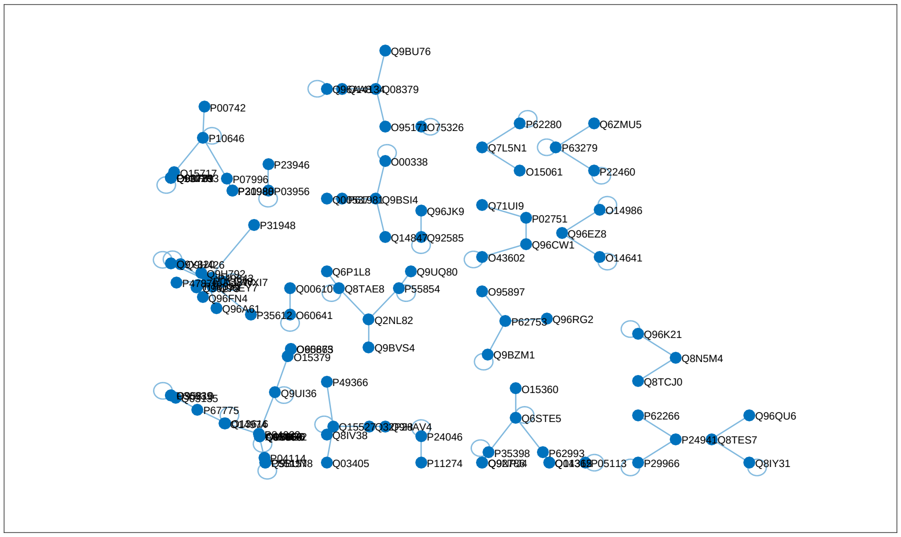


FIGURE A.14: Identified network-biomarkers for subtype 5 distance 3 from seed gene proteins. The maximum size of a subnetwork is 20.

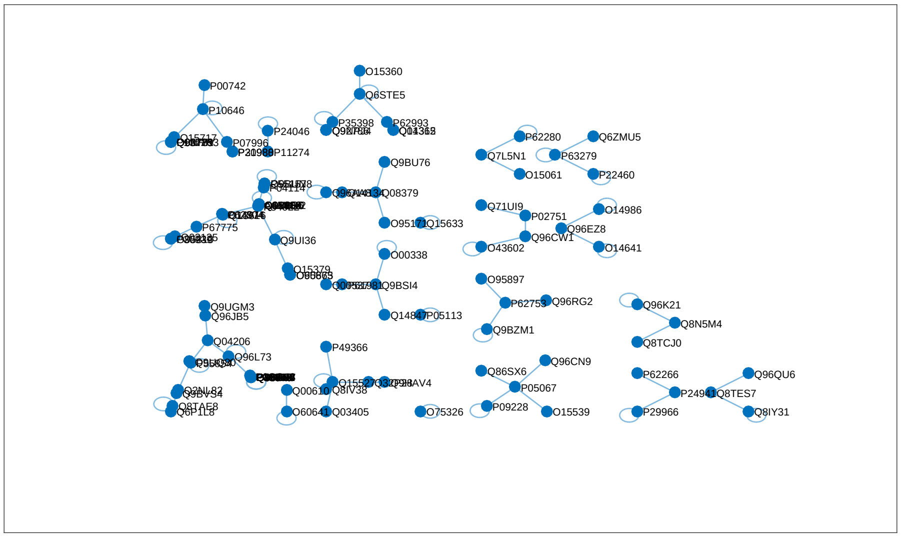


FIGURE A.15: Identified network-biomarkers for subtype 5 distance 4 from seed gene proteins. The maximum size of a subnetwork is 23.

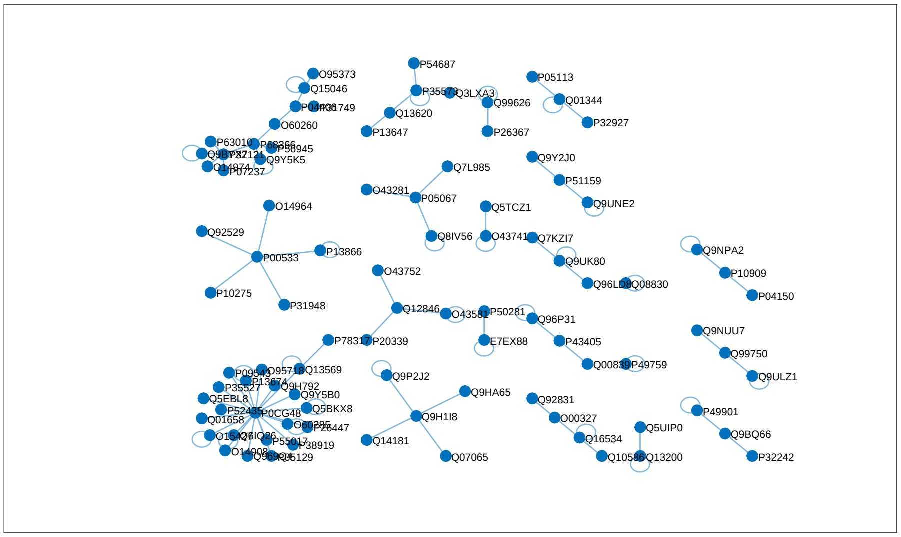


FIGURE A.16: Identified network-biomarkers for subtype 6 distance 2 from seed gene proteins. The maximum size of a subnetwork is 22.

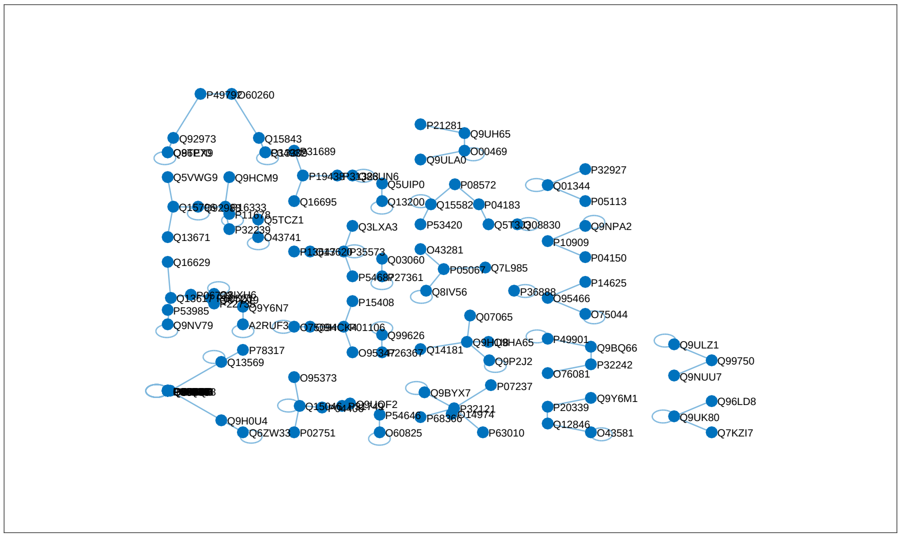


FIGURE A.17: Identified network-biomarkers for subtype 6 distance 3 from seed gene proteins. The maximum size of a subnetwork is 26.

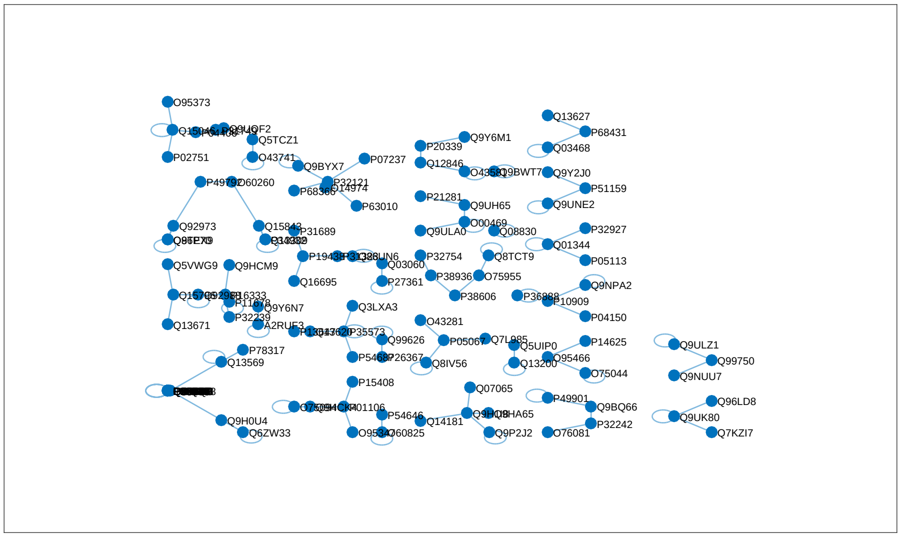


FIGURE A.18: Identified network-biomarkers for subtype 6 distance 4 from seed gene proteins. The maximum size of a subnetwork is 26.

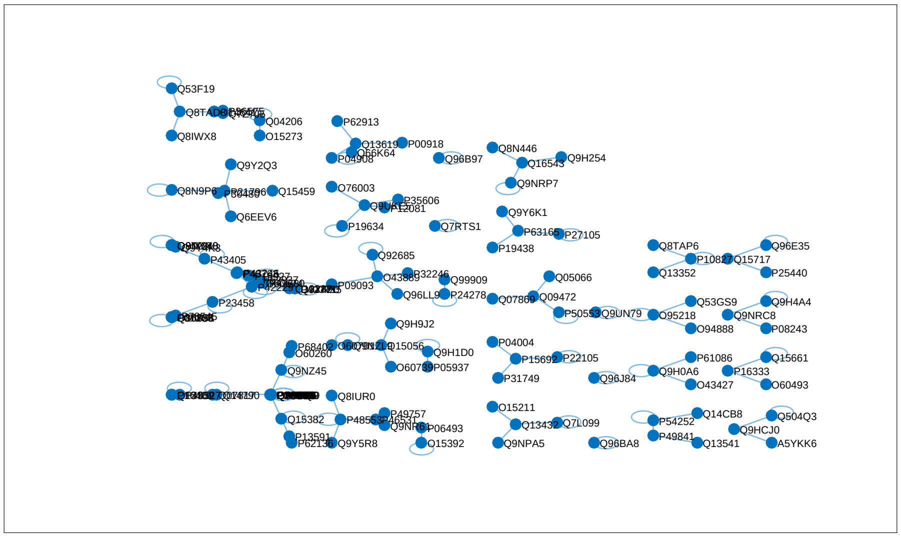


FIGURE A.19: Identified network-biomarkers for subtype 7 distance 2 from seed gene proteins. The maximum size of a subnetwork is 29.

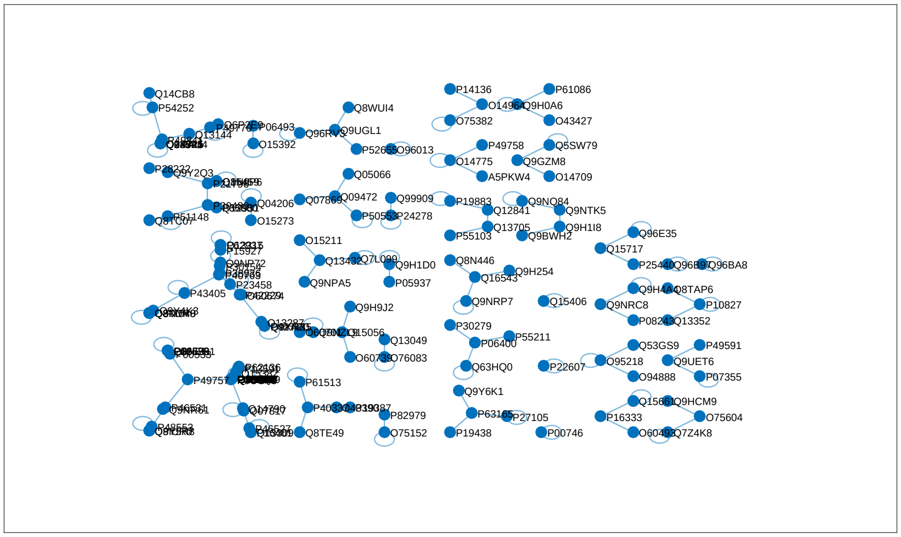


FIGURE A.20: Identified network-biomarkers for subtype 7 distance 3 from seed gene proteins. The maximum size of a subnetwork is 41.

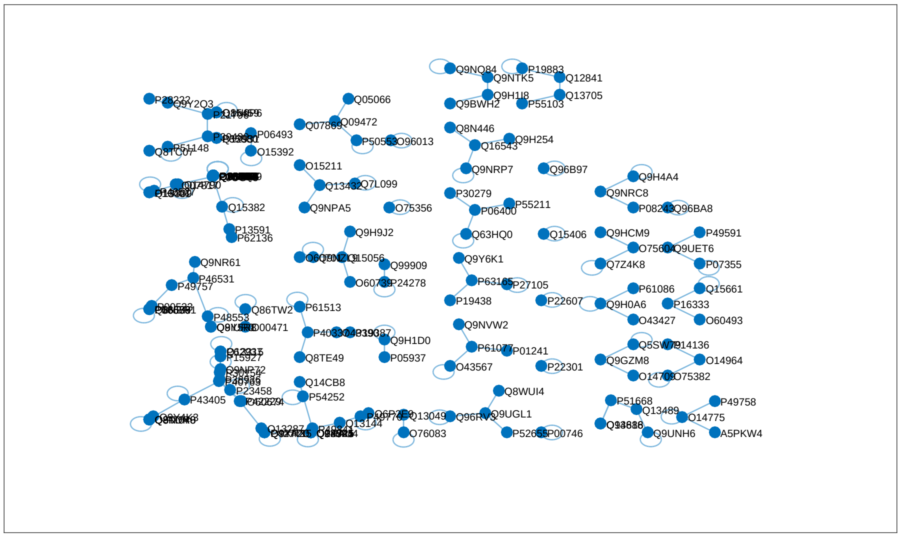


FIGURE A.21: Identified network-biomarkers for subtype 7 distance 4 from seed gene proteins. The maximum size of a subnetwork is 26.

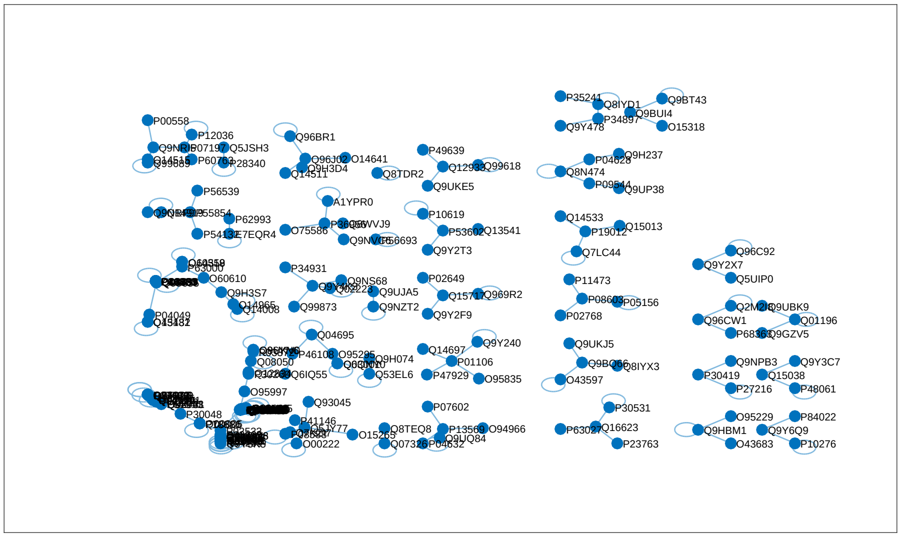


FIGURE A.22: Identified network-biomarkers for subtype 8 distance 2 from seed gene proteins. The maximum size of a subnetwork is 71.

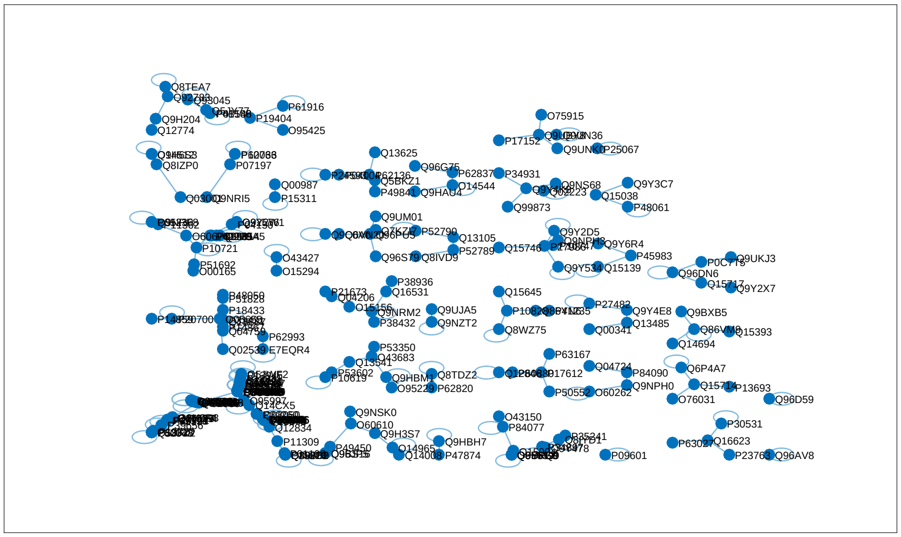


FIGURE A.23: Identified network-biomarkers for subtype 8 distance 3 from seed gene proteins. The maximum size of a subnetwork is 77.

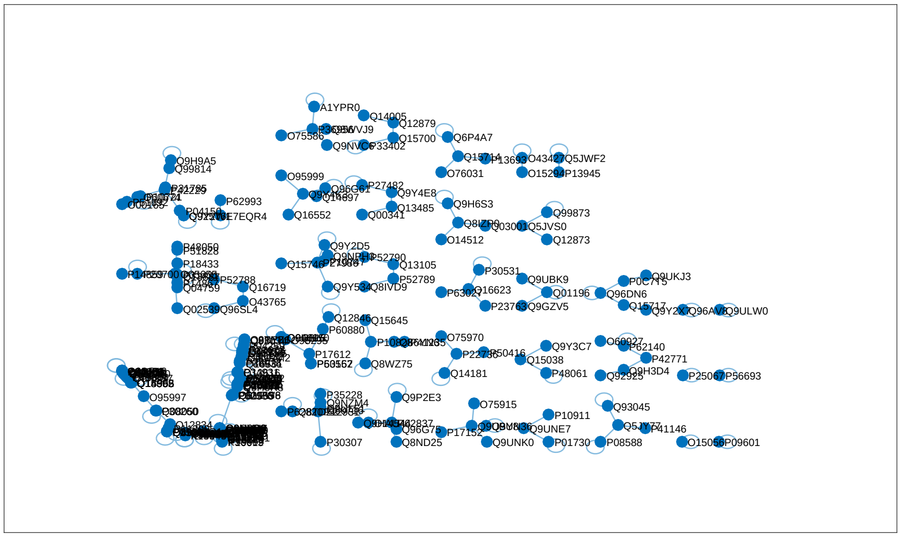


FIGURE A.24: Identified network-biomarkers for subtype 8 distance 4 from seed gene proteins. The maximum size of a subnetwork is 110.

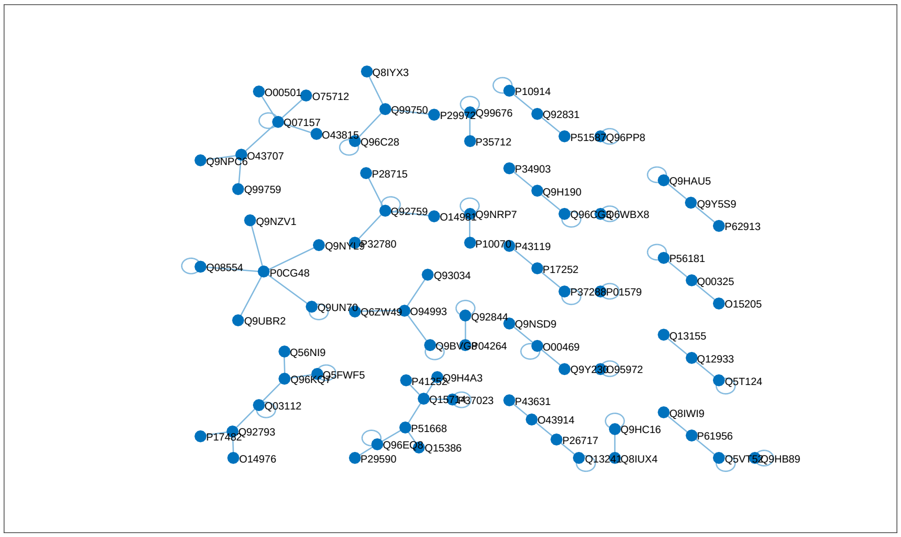


FIGURE A.25: Identified network-biomarkers for subtype 9 distance 2 from seed gene proteins. The maximum size of a subnetwork is 8.

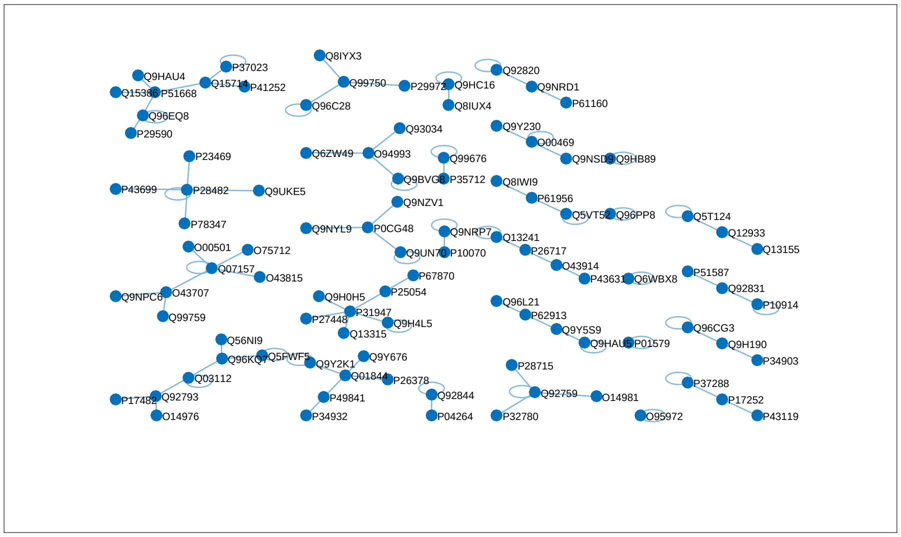


FIGURE A.26: Identified network-biomarkers for subtype 9 distance 3 from seed gene proteins. The maximum size of a subnetwork is 8.

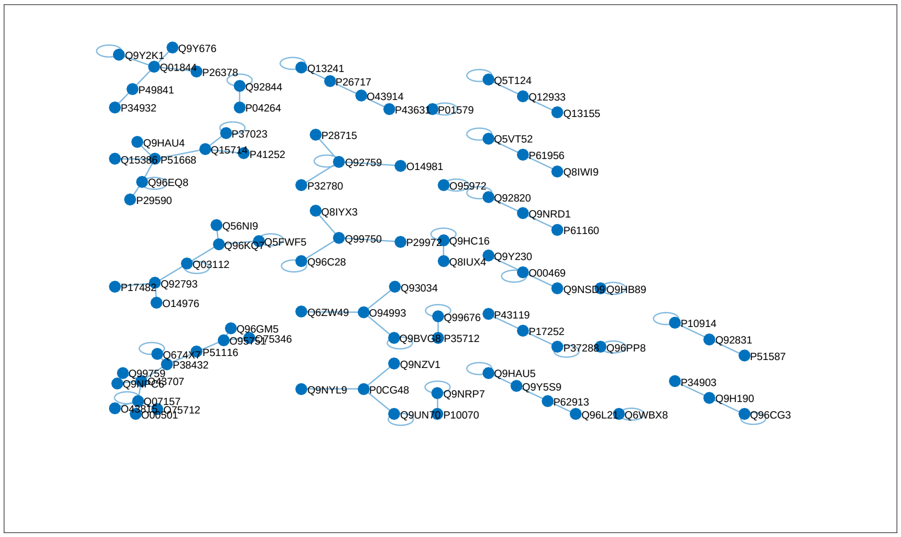


FIGURE A.27: Identified network-biomarkers for subtype 9 distance 4 from seed gene proteins. The maximum size of a subnetwork is 13.

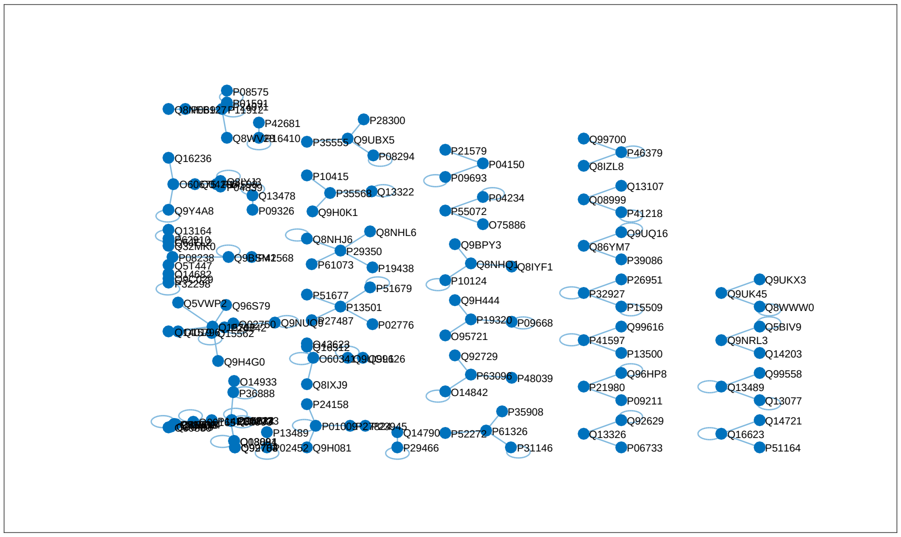


FIGURE A.28: Identified network-biomarkers for subtype 10 distance 2 from seed gene proteins. The maximum size of a subnetwork is 19.

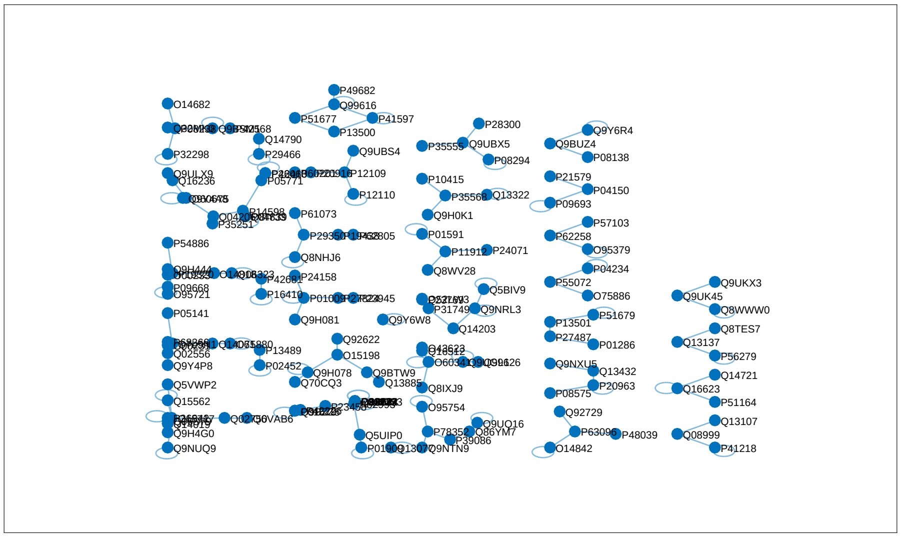


FIGURE A.29: Identified network-biomarkers for subtype 10 distance 3 from seed gene proteins. The maximum size of a subnetwork is 15.

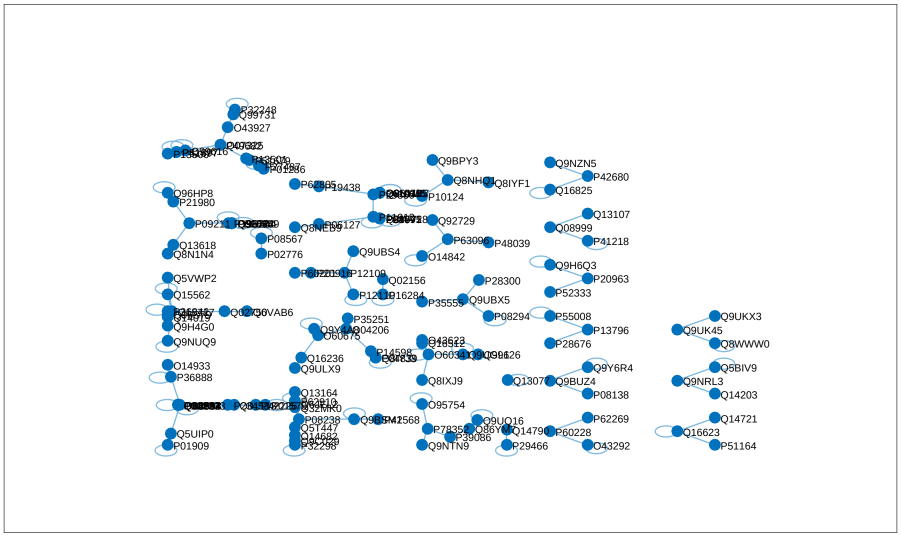


FIGURE A.30: Identified network-biomarkers for subtype 10 distance 4 from seed gene proteins. The maximum size of a subnetwork is 16.

## A.2 Subnetworks for Therapy Specific Survival Prediction

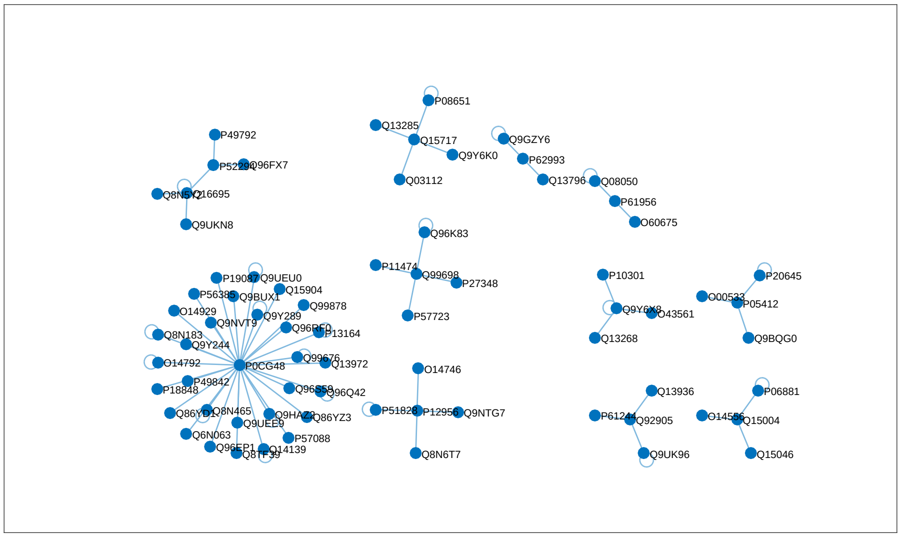


FIGURE A.31: Identified network-biomarkers for chemotherapy received patients with distance 2 from seed gene proteins. The maximum size of a subnetwork is 31.

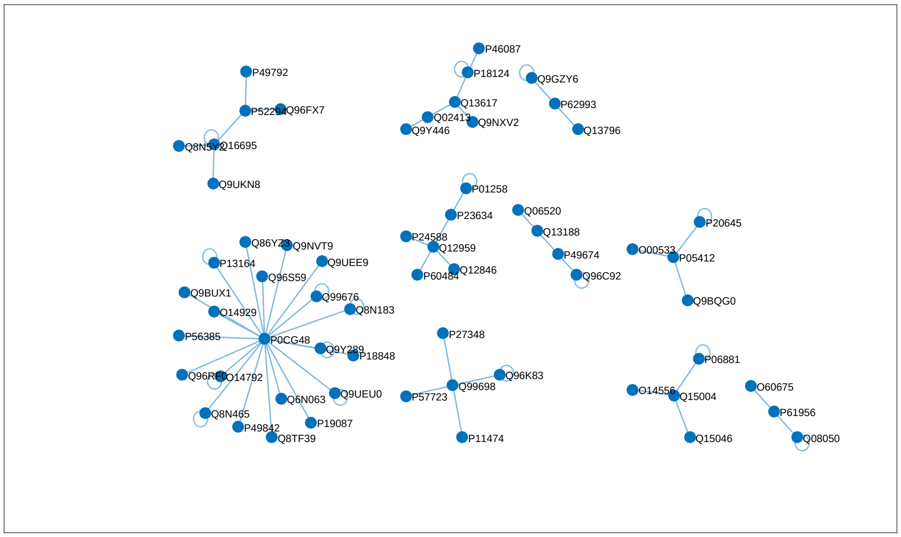


FIGURE A.32: Identified network-biomarkers for chemotherapy received patients with distance 3 from seed gene proteins. The maximum size of a subnetwork is 21.

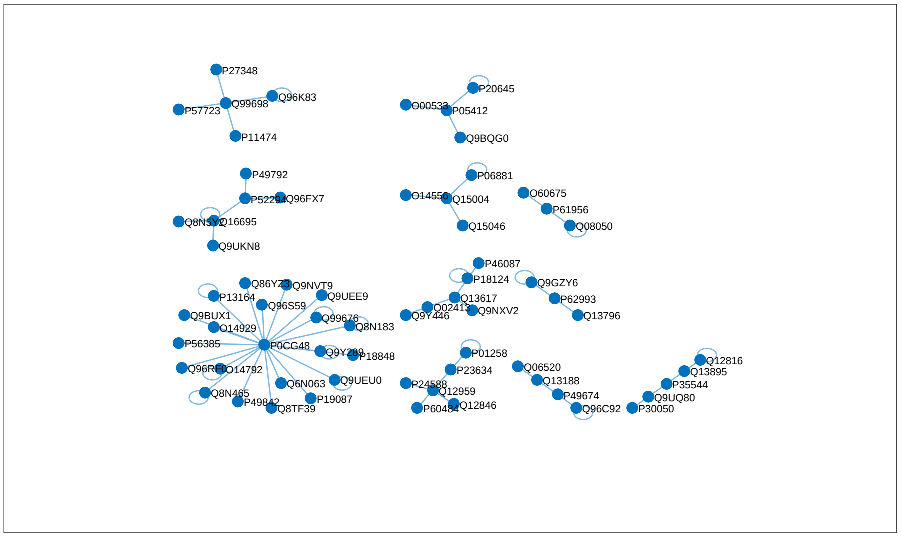


FIGURE A.33: Identified network-biomarkers for chemotherapy received patients with distance 4 from seed gene proteins. The maximum size of a subnetwork is 21.

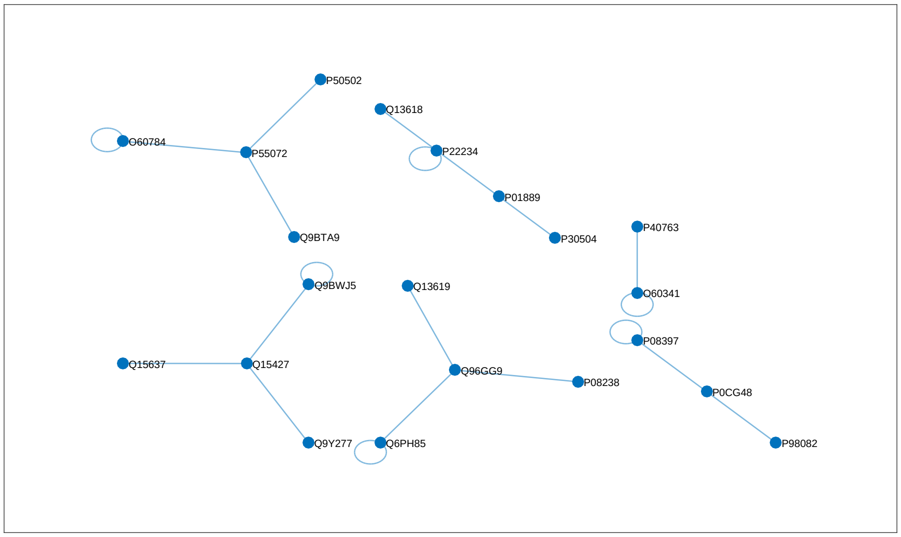


FIGURE A.34: Identified network-biomarkers for chemotherapy and hormone therapy received patients with distance 2 from seed gene proteins. The maximum size of a subnetwork is 4.

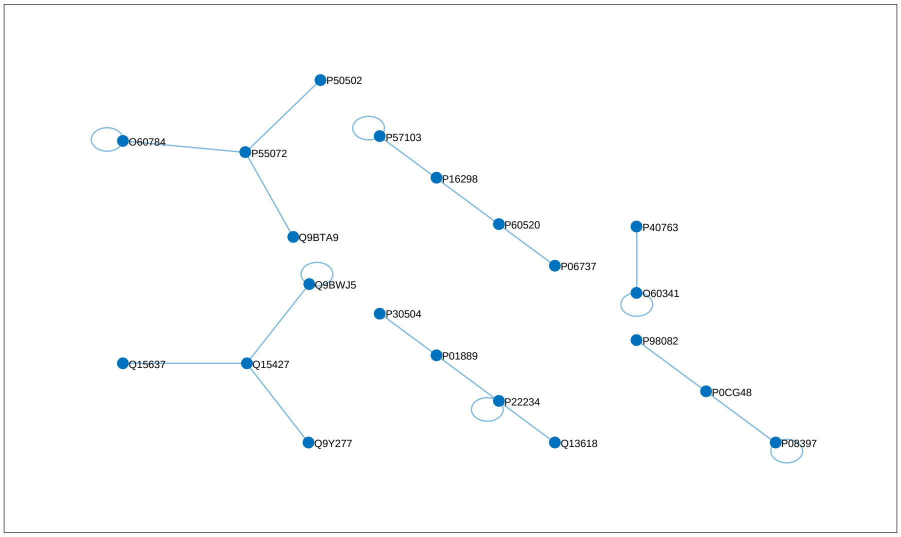


FIGURE A.35: Identified network-biomarkers for chemotherapy and hormone therapy received patients with distance 3 from seed gene proteins. The maximum size of a subnetwork is 4.

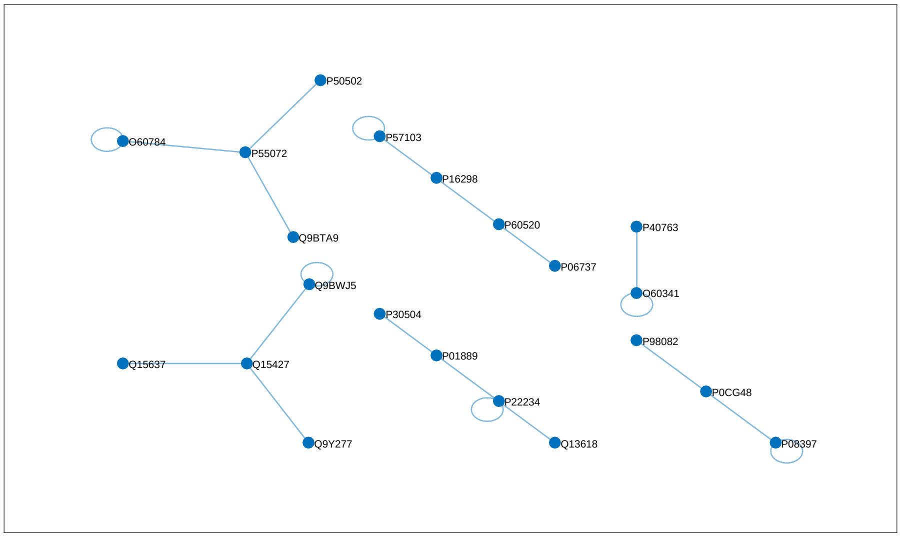


FIGURE A.36: Identified network-biomarkers for chemotherapy and hormone therapy received patients with distance 4 from seed gene proteins. The maximum size of a subnetwork is 4.

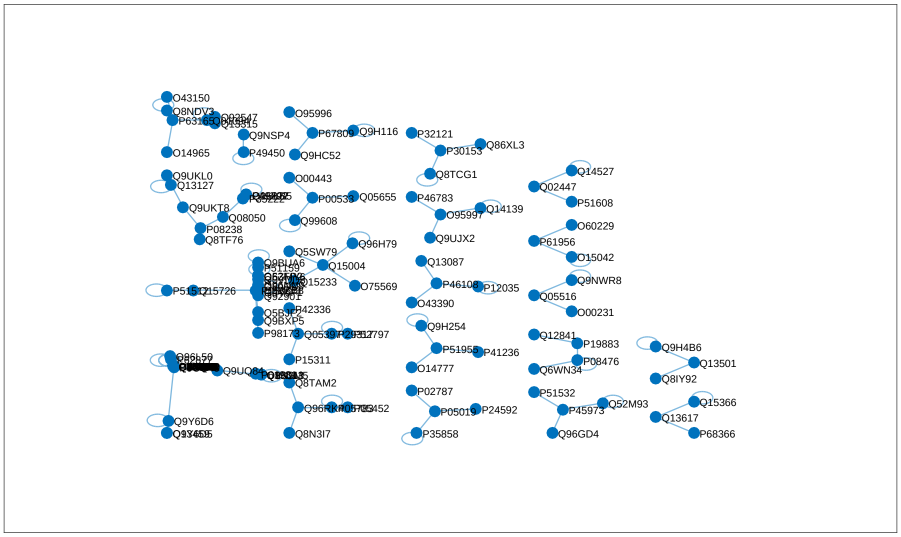


FIGURE A.37: Identified network-biomarkers for chemotherapy, hormone therapy and radiotherapy received patients with distance 2 from seed gene proteins. The maximum size of a subnetwork is 34.

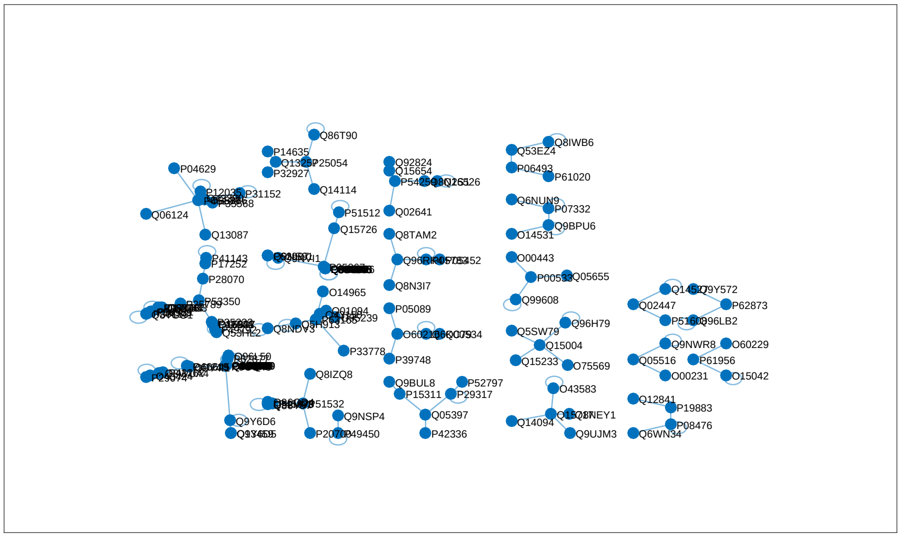


FIGURE A.38: Identified network-biomarkers for chemotherapy, hormone therapy and radiotherapy received patients with distance 3 from seed gene proteins. The maximum size of a subnetwork is 37.

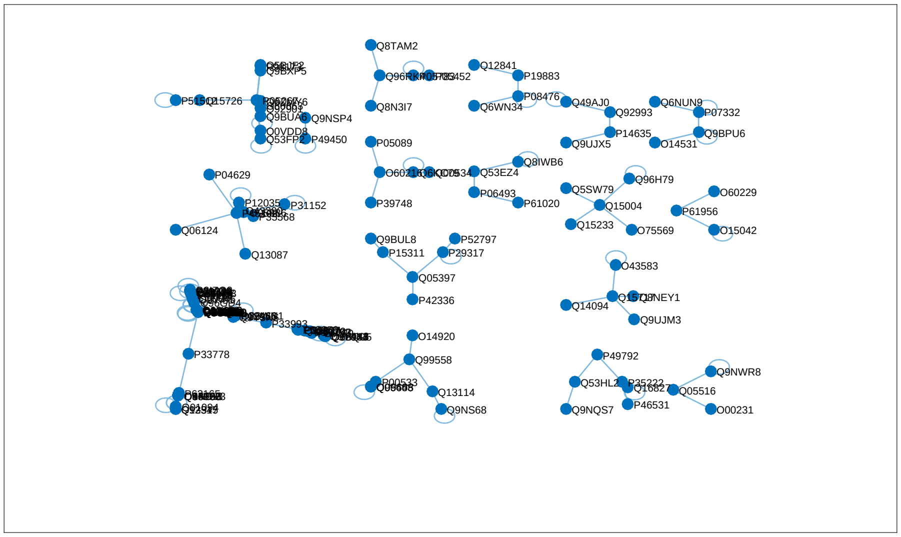


FIGURE A.39: Identified network-biomarkers for chemotherapy, hormone therapy and radiotherapy received patients with distance 4 from seed gene proteins. The maximum size of a subnetwork is 73.

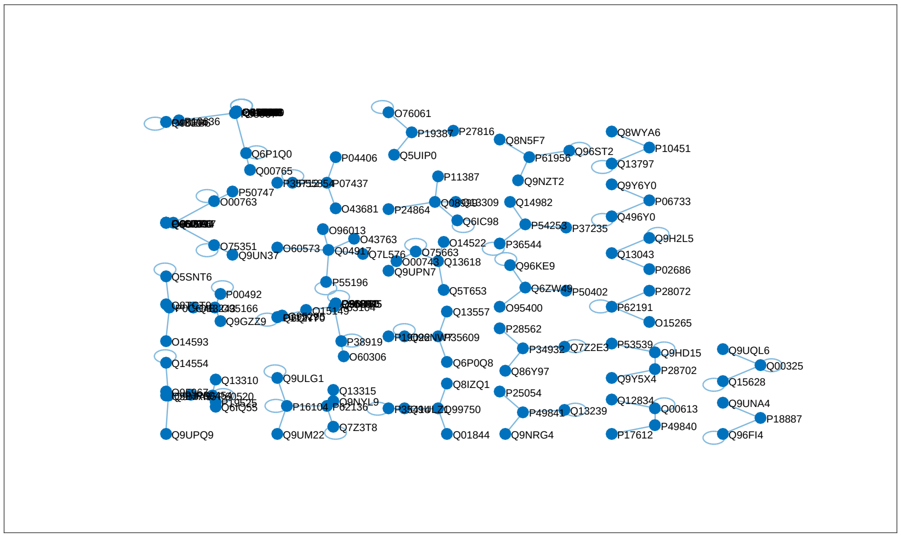


FIGURE A.40: Identified network-biomarkers for chemotherapy and radiotherapy received patients with distance 2 from seed gene proteins. The maximum size of a subnetwork is 23.

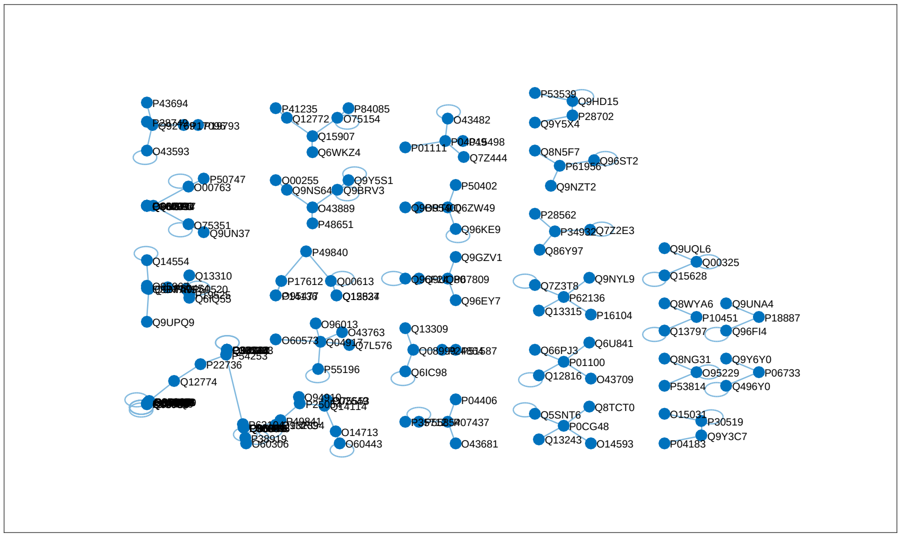


FIGURE A.41: Identified network-biomarkers for chemotherapy and radiotherapy received patients with distance 3 from seed gene proteins. The maximum size of a subnetwork is 44.

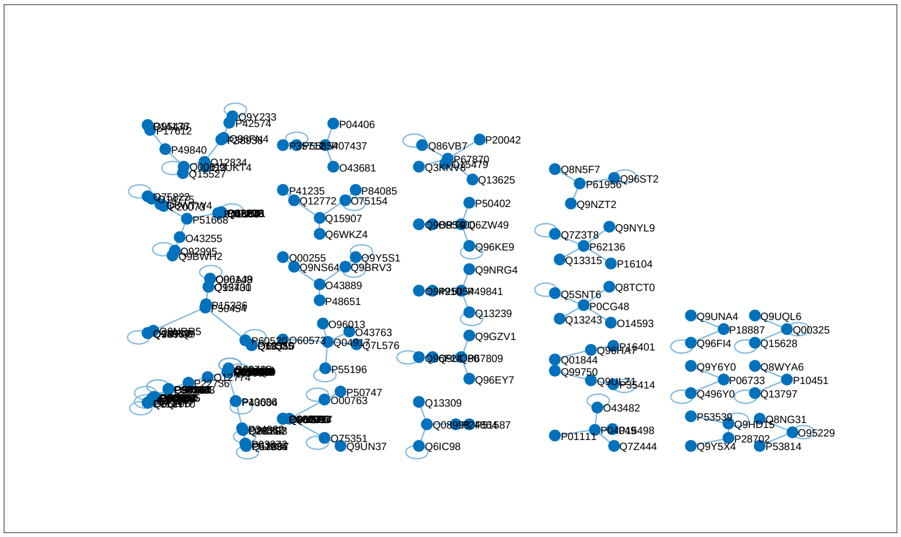


FIGURE A.42: Identified network-biomarkers for chemotherapy and radiotherapy received patients with distance 4 from seed gene proteins. The maximum size of a subnetwork is 55.

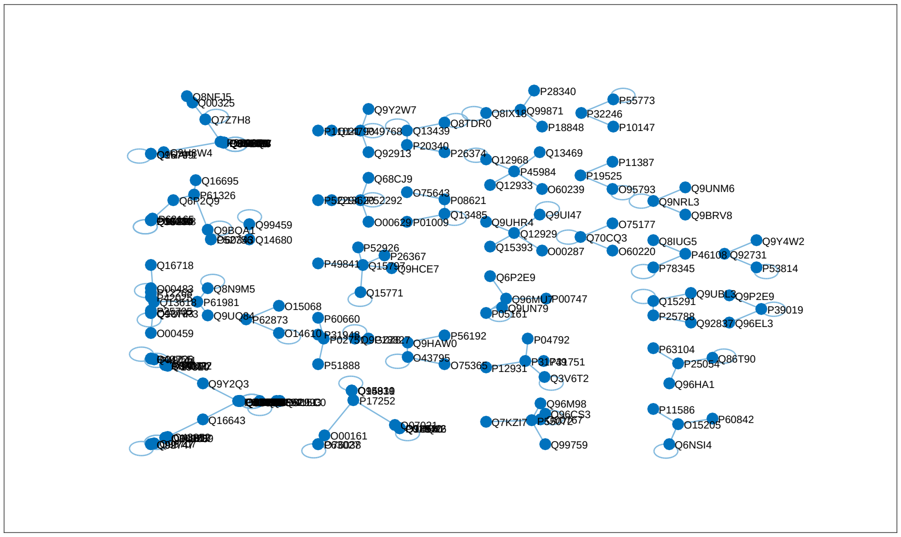


FIGURE A.43: Identified network-biomarkers for hormone therapy received patients with distance 2 from seed gene proteins. The maximum size of a subnetwork is 39.

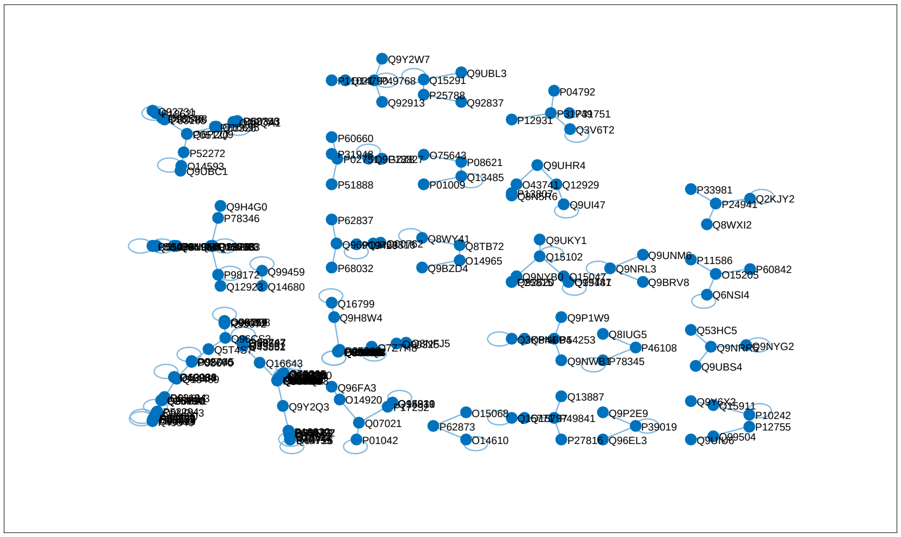


FIGURE A.44: Identified network-biomarkers for hormone therapy received patients with distance 3 from seed gene proteins. The maximum size of a subnetwork is 69.

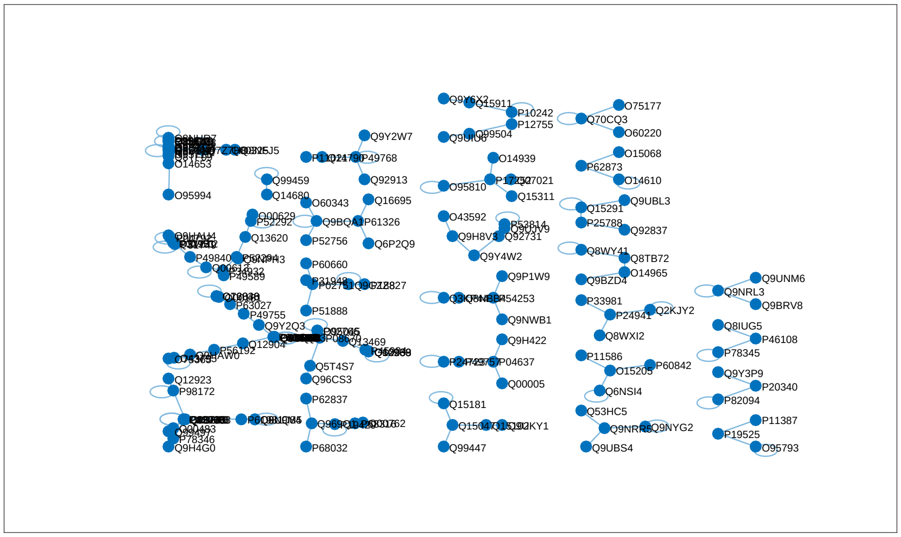


FIGURE A.45: Identified network-biomarkers for hormone therapy received patients with distance 4 from seed gene proteins. The maximum size of a subnetwork is 26.

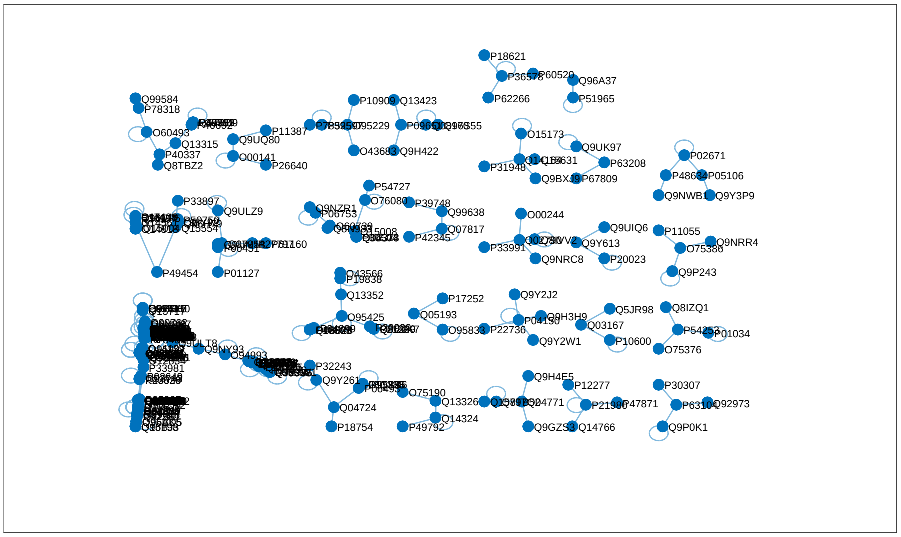


FIGURE A.46: Identified network-biomarkers for hormone therapy and radiotherapy received patients with distance 2 from seed gene proteins. The maximum size of a subnetwork is 165.

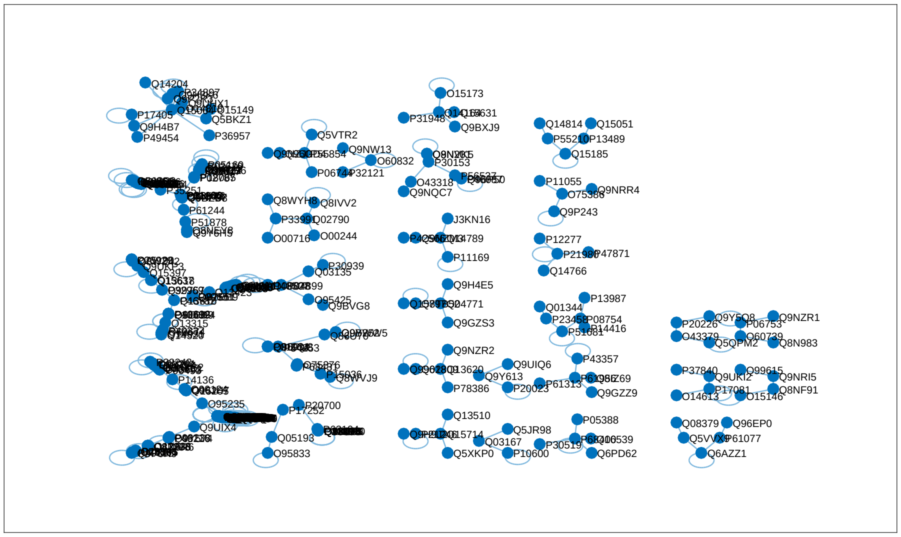


FIGURE A.47: Identified network-biomarkers for hormone therapy and radiotherapy received patients with distance 3 from seed gene proteins. The maximum size of a subnetwork is 135.

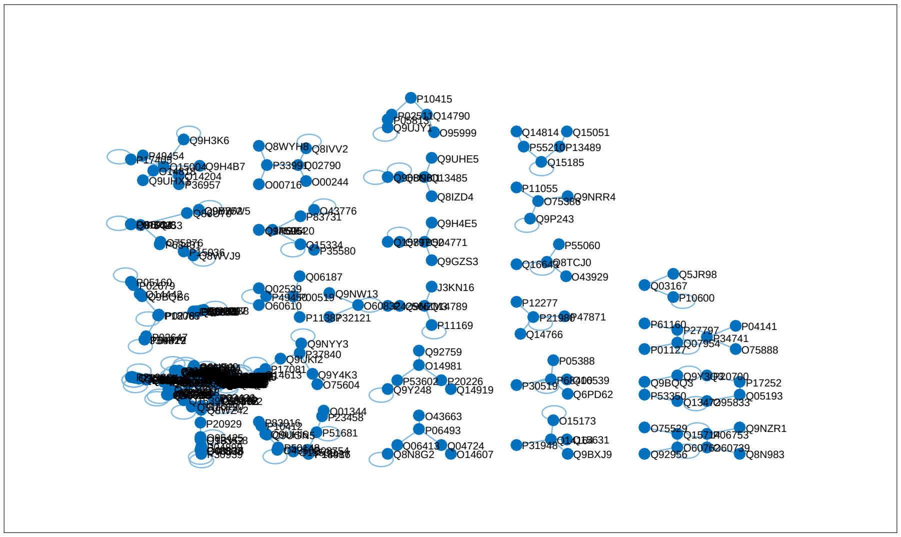


FIGURE A.48: Identified network-biomarkers for hormone therapy and radiotherapy received patients with distance 4 from seed gene proteins. The maximum size of a subnetwork is 198.

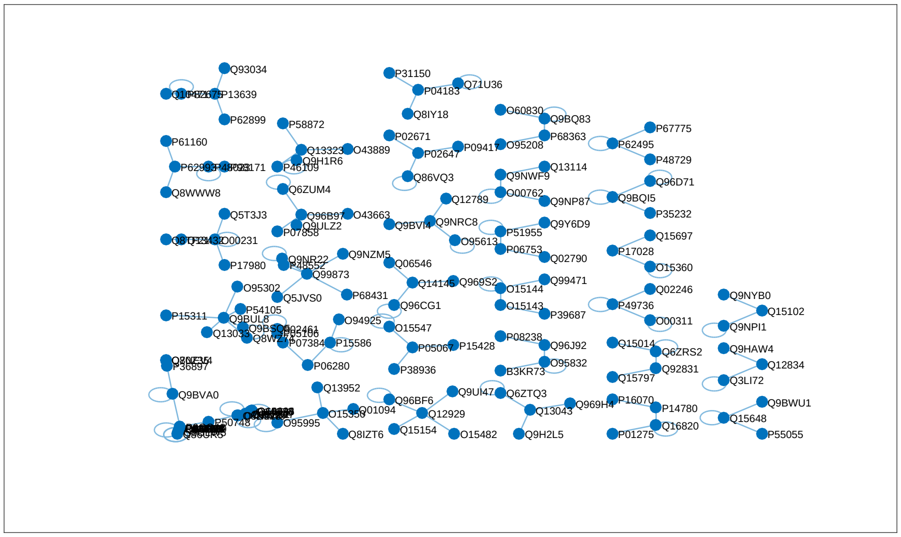


FIGURE A.49: Identified network-biomarkers for radiotherapy received patients with distance 2 from seed gene proteins. The maximum size of a subnetwork is 30.

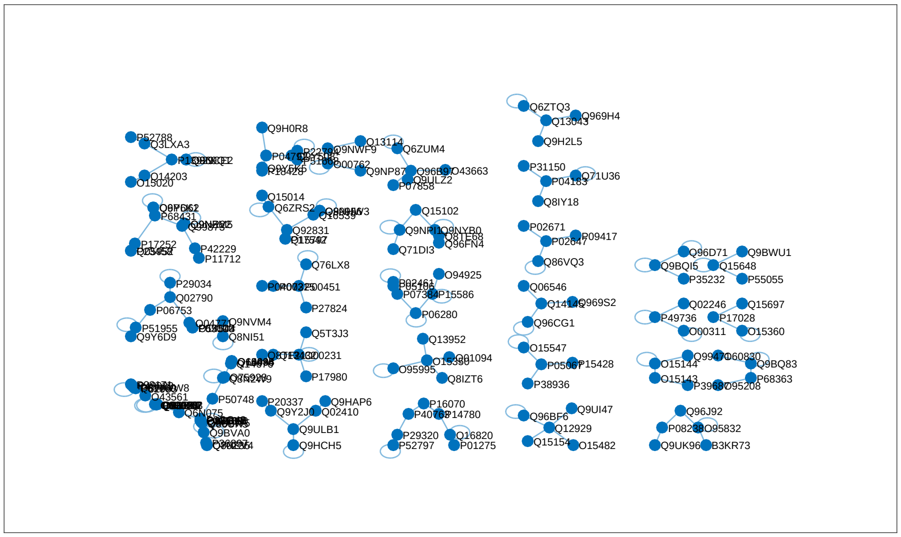


FIGURE A.50: Identified network-biomarkers for radiotherapy received patients with distance 3 from seed gene proteins. The maximum size of a subnetwork is 34.

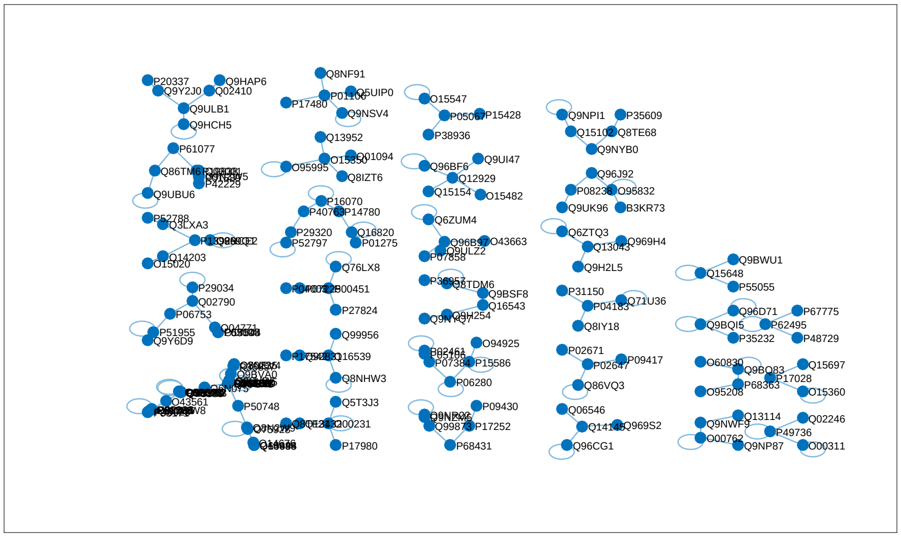


FIGURE A.51: Identified network-biomarkers for radiotherapy received patients with distance 4 from seed gene proteins. The maximum size of a subnetwork is 40.

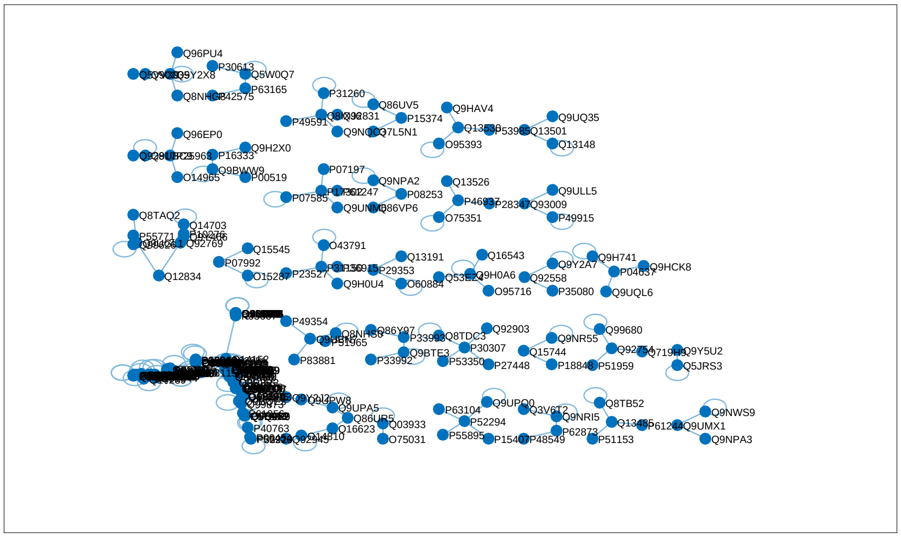


FIGURE A.52: Identified network-biomarkers for no therapy received patients with distance 2 from seed gene proteins. The maximum size of a subnetwork is 149.

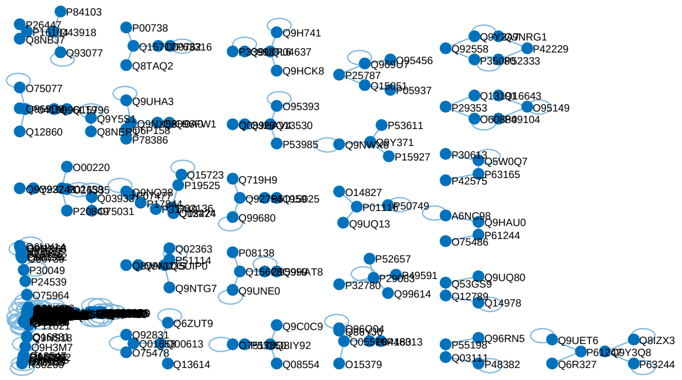


FIGURE A.53: Identified network-biomarkers for no therapy received patients with distance 3 from seed gene proteins. The maximum size of a subnetwork is 184.

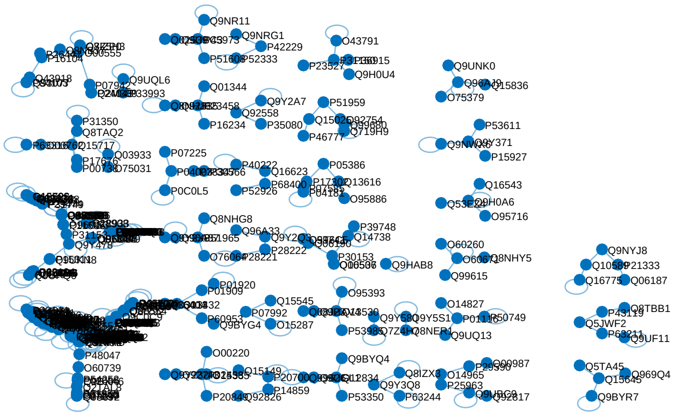


FIGURE A.54: Identified network-biomarkers for no therapy received patients with distance 4 from seed gene proteins. The maximum size of a subnetwork is 143.

# VITA AUCTORIS

| | |
|---|---|
| NAME: | Sheikh Abdullah Al Jubair |
| PLACE OF BIRTH: | Jessore, Khulna, Bangladesh. |
| EDUCATION: | Bachelor of Science in Computer Science, Ahsanullah University of Science and Technology, Dhaka, Bangladesh, 2014.<br>Master of Science in Computer Science, University of Windsor, Windsor, Ontario, Canada, 2017. |