CHALMERS

Centrality in Biological Networks

Master of Science Thesis in Bioinformatics

TRIINU TASA

Chalmers University of Technology
University of Gothenburg
Department of Computer Science and Engineering
Göteborg, Sweden, June 2011

Centrality in Biological Networks

Triinu Tasa

Examiner: Graham Kemp

Chalmers University of Technology
University of Gothenburg
Department of Computer Science and Engineering
SE-412 96 Göteborg
Sweden
Telephone + 46 (0)31-772 1000

Department of Computer Science and Engineering
Göteborg, Sweden June 2011

Centrality in Biological Networks
Master's Thesis in the International Master's Programme in Bioinformatics
Triinu Tasa
Department of Mathematical Sciences
Chalmers University of Technology

Abstract

Centrality analysis has become an important part of biological network studies, notably that of protein-protein interaction networks. It has long been known that the importance of a protein is determined by its connections and relationships to other proteins. In the current work we look into centrality in other kinds of networks as well, notably those based on gene expression data and drug effects on cancer cell lines. The purpose of this project is to show that centrality is useful for the analysis of several different kinds of biological networks. Firstly, we show that the most central genes in the p53 protein interaction network are also the most relevant regarding the network's ability to suppress tumors. Secondly, we look into different types of breast cancer and demonstrate that central genes are among the best discriminators between different classes of data. It is also interesting to see many of the p53 pathway elements coming up among the top central genes. Finally, we apply centrality on cancer treatment data and show how it can be used to identify good drug candidates for different kinds of cancer.

Contents

Acknowledgements

Most of the research for this thesis was done in collaboration with the Newcastle Bioinformatics Initiative group at the University of Newcastle, Australia. I would like to thank the entire group for their continous support and guidance throughout the entire project. Especially I would like to thank Prof. Pablo Moscato and Dr. Luke Mathieson, and also Dr. Mario Inostroza-Ponta and Dr. Alexandre Mendes for their help with the computational aspects of the work. Additionally I would like to thank Dr. Graham Kemp for reviewing this work and providing me with valuable feedback.

1 Introduction

Until recently the importance of a gene was determined by its individual actions as catalysts, signalling molecules, or building blocks of cells. It is increasingly clear though, that most of its biological characteristics are determined by interactions with other constituents of the cell, such as proteins, DNA, RNA, and small molecules. Therefore, a lot of time and effort is put into constructing different kinds of networks based on these interactions.

Various types of interaction networks emerge from the sum of these interactions, including protein-protein interaction, metabolic, signalling and transcription-regulatory networks. It is also possible to construct networks based on gene expression data where genes with similar expression profiles are linked together. None of these networks are independent, instead they form a ”network of networks” that is responsible for the behavior of the cell. In the current project we look into protein-protein and gene expression networks, and last but not least, a network constructed from drug reaction data on different cancer cell lines.

Networks are usually represented in the form of graphs. Informally speaking, a graph is a set of vertices (also referred to as nodes) and edges (links) connecting them. In protein-protein interaction networks proteins are denoted by nodes, and a link represents a mutual binding relationship: if protein A binds to protein B, then protein B also binds to protein A. In gene expression networks the genes are represented as nodes, and connections between genes symbolize similarity in expression. Finally, in drugs vs. cancer cell-lines we denote different drugs with nodes and the connections between the drugs represent their similar inhibition properties on cancer cell lines.

The purpose of this work from the computational point of view is to find the most important vertices in the graph. The importance of a vertex is determined by the level of damage (the connectivity of the whole graph diminishes greatly) that the removal causes. It has been shown in several works [21, 12] that the genes representing the most important vertices are often also the most important ones in their biological networks. In protein-protein interaction networks a cancellation or mutation in any of these genes often leads to serious consequences and might even be lethal for the cell. The significance of the central genes of gene expression and drug vs. cell line networks will be studied further in chapter 3.

Chapter 2 discusses all the computational issues - gives an overview of the relevant graph theory, discusses the algorithms for finding the vital vertices in a graph and explains the most important aspects of the developed software. The second part of the work puts the previously developed software into use. Firstly, in section 3.1 we examine the p53 protein interaction network and look into its most central proteins and their significance in the p53 pathway. In section 3.2 we analyze the gene expression data on two different types of breast cancer. We cluster the genes by expression similarities and analyze the most central genes in some of the most interesting clusters. We also study where the central proteins from the p53 pathway are placed in this gene expression network. Finally in section 3.3 we construct a network based on drug inhibition properties on different cancer cell lines and try to employ centrality for locating the best compounds for treating breast cancer.

2 Theoretical background

2.1 Scale-free networks

For decades graph theory was focused on either regular or completely random graphs. However, neither model is suitable for describing most real-life networks like social networks, internet and biological networks (protein-protein interaction, metabolic, transcription-regulatory and signalling networks). Only recently a new graph model was introduced which describes all of these networks - the model of scale-free networks [3].

It was first noted that for many real-life networks the number of nodes with a given degree follows a power law. That is, the probability that a chosen node has exactly k links follows $P(k) \sim k^{-\gamma}$, where γ is the degree exponent with its value for most networks being between 2 and 3 [3]. Most networks only have a few nodes with a large number of links (often called hubs) whereas most nodes only have a few. Such networks are called scale-free [3].

Scale-free networks are amazingly robust against accidental failures - even if 80% of randomly selected nodes fail, the remaining 20% still form a compact cluster with a path connecting any two nodes [1]. On the other hand, a lot of damage might be caused if a few key hubs are knocked out.

An important feature of every network is its average path length. Path length is the number of links we need to pass through to travel between two nodes. Average path length represents the average over the shortest paths between all pairs of nodes and offers a measure of a network's overall navigability. A common feature of all complex networks is their small average path length (known as the ”small world effect”) [4]. Scale-free networks are ”ultra small” [5, 6]. Path length in scale-free networks with degree exponents $2 < \gamma < 3$ is even smaller, with the average path length following $l \sim \log \log N$ [5, 6], which is significantly shorter than $logN$ that characterizes random small-world networks.

It has been shown in several organisms now that most networks, including protein-protein interaction networks, within the cell approximate a scale-free topology [21, 13, 24]. The reason for that lies in the evolutionary history of biology. For example, protein-protein interaction networks are believed to be scale-free because of gene duplication

[30]. Duplicated genes produce identical proteins that interact with the same protein partners. Therefore, each protein that is in contact with a duplicated protein gains an extra link. The scale-free model predicts that the nodes that appeared early in the history of the network are the most connected ones [3].

The ultimate description of cellular networks requires some extra information in addition to the interaction schema. The intensity and temporal aspects of interactions must also be considered, as some are more active than others and some are active only during a certain period. So far we only have little information about the temporal aspects of various cellular interactions, but our knowledge of intensities is improving. In protein-protein interaction networks we often have information about the intensity of a relationship, which is represented in the graph as the weight of the link.

This section is, to a large extent, based on the excellent review by Barabási and Oltvai [4], which gives an overview of different graph models and their usability in biological networks.

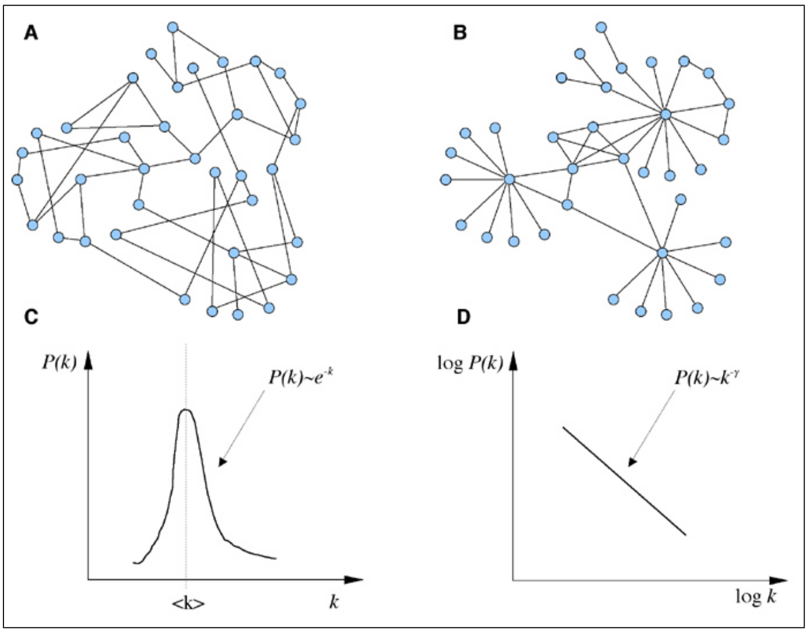


Figure 2.1: Examples of A) random and B) scale-free network (from [20])

In figure 2.1 both networks have 36 nodes and 44 links. However the organization of connections makes the difference. In the random network, majority of nodes have 2 or 3 connections, making it follow Poisson distribution (C). Scale-free networks with a relatively small number of hubs follow power law $P(k)$ (D), which is defined as the probability that a randomly chosen node in the network has exactly k links.

2.2 Average path length

Path length is the distance between two vertices in the graph. Average path length of a graph is the average of the minimal path lengths between all vertices in the graph. The mathematical formula for calculating the average path length is given in equation 2.1:

$$APL_G = \frac{\sum_{i=1}^{n} \sum_{j=i}^{n} l_{ij}}{n^2}, \tag{2.1}$$

where G - graph G,
n - numer of vertices,
l_{ij} - minimal path length between vertices i and j

In the current work however we will not need the average path length of the whole graph but we will need it for each vertex separately. The average path length for each vertex i is calculated as the average of all minimal path lengths l_{ij}, where $j \in V \backslash \{i\}$ and V is the set of all vertices. The corresponding formula is given in equation 2.2.

$$APL_i = \frac{\sum_{j=1}^{i-1} l_{ij} + \sum_{j=i+1}^{n} l_{ij}}{n-1} \tag{2.2}$$

Various algorithms can be used for calculating the average path lengths. I compared two main algorithms, Dijkstra (Algorithm 1) and Floyd-Warshall (Algorithm 2) algorithms. Dijkstra in essence is intended for finding the shortest path between two given vertices. Floyd-Warshall algorithm in contrast calculates all shortest paths in the graph. However, adapting Dijkstra to calculate all shortest paths resulted in an algorithm that works at least as fast as Floyd-Warshall or even faster for larger data amounts.

The Dijkstra algorithm (Algorithm 1) works by memorizing for each vertex v the cost $d[v]$ of the shortest path found so far between s and v. Initially, this value is 0 for the source vertex s ($d[s] = 0$), and infinity for all other vertices, which means that we do not know any path leading to those vertices ($d[v] =?$ for every v in V, except s). The notation $w(u, v)$ represents the cost of traversing the edge between the vertices u and v. $Extract\_Min(Q)$ retrieves the vertex with the shortest path from vertex s, which initially is s itself. When the algorithm finishes, $d[v]$ will be the cost of the shortest path from s to v or infinity, if no such path exists [47].

The Floyd-Warshall algorithm (2) is a brute-force algorithm calculating all distances between all vertices and memorizing the shortest paths.

Algorithm 1 Dijkstra(G, t, s)

```
for all vertex v in V[G] do
  d[v] := infinity
  previous[v] := undefined
end for
d[s] := 0 //Distance from s to s
S := empty set //Set of all vertices
Q := V[G]
while Q is not an empty set do
  u := Extract_Min(Q)
  S := S union {u}
  for all edge (u,v) outgoing from u do
    if d[u] + w(u,v) < d[v] then
      d[v] := d[u] + w(u,v)
      previous[v] := u
    end if
  end for
end while
```

2.3 Clustering

Throughout this work we are using clustering for grouping together similar data and for finding the nearest neighbors of the node of interest. The algorithm we opted for combines the Minimum Spanning Tree and k-Nearest Neighbors algorithms and is hence abbreviated as MSTkNN. It constructs a disconnected graph by computing the intersection of the outputs of the two algorithms mentioned above. The algorithm was first presented by González-Barrios and Quiroz [14], but in this work we are using a modification of it by Inostroza-Ponta[17].

A spanning tree of a connected, undirected graph is a subgraph which is a tree and connects all the vertices together. There can be many different spanning trees, and a **minimum spanning tree** is the one with the smallest weight. The total weight of a spanning tree is calculated by summing up all individual weights of the edges [56].

k-nearest neighbor algorithm is a simple classification method where an object is assigned to a class that the majority of its nearest neighbors belong to [57].

Algorithm 2 Floyd-Warshall($int[1..n, 1..n]$ graph)

```
//Initializing the distance matrix dist[i][j]
for all i = 1 to n do
  for all j = 1 to n do
    if i == j then
      dist[i][j] := 0
    else if exists distance from i to j then
      dist[i][j] := distance_i_j
    else
      dist[i][j] := infinity
    end if
  end for
end for
//Main loop of the algorithm
for all k = 1 to n do
  for all i = 1 to n do
    for all j = 1 to n do
      if dist[i][j] > dist[i][k] + dist[k][j] then
        dist[i][j] = dist[i][k] + dist[k][j]
      end if
    end for
  end for
end for
```

2.3.1 Distance metrics

There are different ways to calculate the weights of the minimum spanning tree. In the current work we compared 3 different methods: Euclidean distance, Pearson correlation and Spearman correlation.

Euclidean distance is calculated between two points p and q using the following formula [60]:

$$d(p,q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2} \tag{2.3}$$

When working with gene expression data we are using expression values for p and q, and n is the number of probes. Euclidean distance works reasonably well when data is normalised and values in rows are comparable. In gene expression data that hasn't been normalised it cannot be used though, as values of different probes can differ by thousands of times. In such cases it is better to use correlation.

Pearson correlation coefficient is widely used in the sciences as a measure of the strength of linear dependence between two variables. Pearson's correlation coefficient between two variables is defined as the covariance of the two variables divided by the product of their standard deviations [58]:

$$\rho = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_x \sigma_Y} \tag{2.4}$$

The above formula defines the population correlation coefficient. Substituting estimates of the covariances and variances based on a sample gives the sample correlation coefficient, commonly denoted r:

$$r = \frac{1}{n-1}\sum_{i=1}^{n}[(\frac{X_i - \bar{X}}{s_X})(\frac{Y_i - \bar{Y}}{s_Y})] \tag{2.5}$$

where $\frac{X_i - \bar{X}}{s_X}$, $\bar{X}$, and s_X are the standard score, sample mean and sample standard deviation respectively [58].

Another correlation coefficient that we use is the **Spearman's rank correlation coefficient**. It is a non-parametric measure of statistical dependence between two variables. It assesses how well the relationship between two variables can be described using a monotonic function. If there are no repeated data values, a perfect Spearman correlation of +1 or 1 occurs when each of the variables is a perfect monotone function of the other [59].

The formula for calculating the Spearman correlation is the same as Pearson's correlation coefficient with the exception of using ranks instead of values for x_i and y_i:

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} \tag{2.6}$$

Pearson correlation only gives a perfect value when X and Y are related by a linear function. In contrast Spearman correlation gives a perfect value when X and Y are related by any monotonic function, which is why it is often described as being nonparametric [59].

3 Tests

3.1 The p53 Protein Interaction Network

3.1.1 Introduction

The p53 pathway has caught the attention of hundreds of scientists because of its role in apoptosis, cellular senescence (aging) and cell cycle arrest [23]. It is also involved in preventing DNA damage and repairing already damaged DNA [23]. One of the most central players in the p53 pathway is the p53 protein.

There are many reasons for DNA damage like UV irradiation, duplication errors, reaction with oxidative free radicals and many more. Each type of DNA damage is detected and fixed by a different set of proteins, but they are all reported to the p53 protein and its pathway [23]. The p53 pathway acts as a supervisor and removes the cells with DNA errors [23].

Considering the function of the p53 network it is only natural that any major mutation in the network's central genes can have severe consequences. In fact it has been observed that the p53 gene is mutated about 50% of the time in a wide variety of cancers, and at other times the proteins that frequently interact with p53 are mutated [39].

In this thesis we concentrate on a simplified version of the p53 pathway - the p53 protein interaction network. Protein-protein interactions occur when two or more proteins bind together. The graph of the network is shown below (3.1). As the p53 network is a scale-free network, it is amazingly robust to the knockout of most genes whereas mutations in some of the most connected or central ones often result in the development of cancer. Next we explore the most central genes in more detail.

3.1.2 The p53 gene

In order to get a better grip of the p53 network, we first calculated the average path lengths of all the genes in the network and then sorted the results accordingly. The 25 most central genes are given below with average path length in the brackets :

- 1. p53 (1.9)

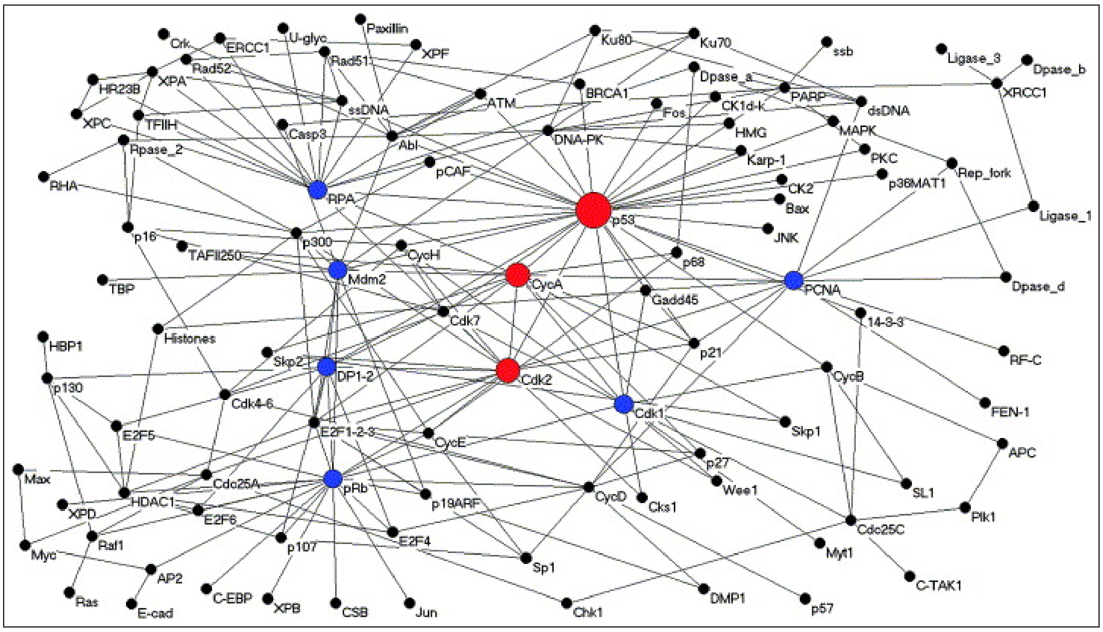


Figure 3.1: The p53 network [7]

- 2. Cdk2 (2.1)
- 3. CycA (2.2)
- 4-7. Cdk1, Mdm2, DP1-2, pRb (2.3)
- 8-12. PCNA, RPA, DNA-PK, p21, p300 (2.4)
- 13-15. E2F1-2-3, CycH, Cdk7 (2.5)
- 16-17. Abl, Gadd45 (2.6)
- 18-22. CycB, CycD, CycE, PARP, ATM (2.7)
- 23-25. ssDNA, Cdc25A, 14-3-3 (2.8)

The most important gene in the p53 network is the p53 gene itself. According to average path length calculations it is also the most central. The protein product of the p53 gene responds to stress signals caused by DNA damage and acts upon it. Whenever the p53 protein is informed about DNA damage, the concentration of the protein is increased [23]. As p53 also regulates the transcription genes like MDM-2, p21, 14-3-3 sigma and GADD45, the concentration rates of those also increase [23]. Most of the genes regulated by p53 are involved in apoptosis.

The most interesting feature of p53 for us is it's activation when certain tumor suppressor genes get inactivated because of a mutation. Such genes include oncogenes like mys and Ras, the APC tumor suppressor and the retinoblastoma protein pRb. Mutations in the retinoblastoma proteins can cause the development of the cancer of retina (in the eye), which usually occurs in children of less than 5 years of age. pRb acts by regulating the E2F-1 transcription factor, which controls cell proliferation and apoptosis. Mutations in the APC gene often contribute to the development of colon cancer. Oncogenes are genes that code for a protein that is believed to cause cancer. For example, when myc is specifically mutated, or overexpressed, it increases cell proliferation [49]. Proteins in the Ras family control such processes as cytoskeletal integrity, proliferation, cell adhesion, apoptosis, and cell migration. Ras proteins are often deregulated in cancers, leading to increased invasion and metastasis, and decreased apoptosis [50].

Cdk2 is a catalytic subunit of the cyclin-dependent kinase complex, whose activity is restricted to the G1-S phase of the cell cycle, and is essential for the G1/S transition. This protein associates with and is regulated by the regulatory subunits of the complex including cyclin E or A. Cyclin E binds G1 phase Cdk2, which is required for the transition from G1 to S phase while binding with Cyclin A (CycA) is required to progress through the S phase.

3.1.3 Other central genes

In this section we will discuss some of the genes from the listing above that have not yet been mentioned.

As can be seen from the graph in figure 3.1, one of the most central genes is DP1-2. It is a cell cycle regulatory transcription factor that forms a functional heterodimer with E2F1. The dimer can in turn bind to MDM-2, which is involved in cell cycle arrest. In addition, it is known that DP1-2 expression is strongly inhibited by p53 at the level of transcription. Inhibition of DP1 transcription has implications in one of the several possible mechanisms through which p53 induces cell cycle arrest [15].

PCNA or Proliferating Cell Nuclear Antigen is a protein that is transcriptionally activated by p53 and is important for both DNA synthesis and DNA repair.

Replication protein A (RPA) is required for both DNA replication and nucleotide excision repair. In normal state RPA binds to p53, but it has been shown that in the case where the cell is radiated with UV, the ability of RPA binding to p53 is greatly reduced. In conclusion it has been proposed that RPA may participate in the coordination of DNA repair by releasing p53 when sensing UV damage. When released, p53 can act to repair any damage.

ATM and DNA-PK are protein kinases that in addition to responding to DNA damage are also involved in controlling genome stability and cell cycle progression [22].

Insufficient levels of ATM in humans can cause neurodegeneration, immunodeficiency, genome instability and cancer predisposition [22]. The deficiency of DNA-PK in mice leads to severe immunodeficiency [22]. ATM and DNA-PK phosphorylate p53 which initiates activation of the DNA damage checkpoint, leading to cell cycle arrest, DNA repair or apoptosis.

p300 is a transcriptional co-activating protein. The members of p300 family are transcriptional adaptors for p53, modulating its checkpoint function in the G1 phase of the cell cycle and its induction of apoptosis [26].

When Abl oncogene is translocated within the bcr (breakpoint cluster region) gene it activates a tyrosine kinase which allows the cells to proliferate without being regulated, leading to chronic leukemia [51].

PARP (Poly ADP-ribose polymerase) is a protein involved in a number of cellular processes involving mainly DNA repair and programmed cell death. It has been suggested that PARP-1, a protein in the PARP family, participates in the p53 response following irradiation [40].

CDC25A is a member of the CDC25 family of phosphatases. CDC25A is required for progression from G1 to the S phase of the cell cycle [52]. Overexpression of Cdc25A phosphatase is often observed in cancer and results in poor prognosis. Cdc25A mainly dephosphorylates and thereby activates CDK2 and thus induces progression in the cell cycle from G1 to S phase. p53 downregulates expression from the Cdc25A gene [36].

Now that some of the most important genes in the p53 network have been shortly described, we continue with studying their behaviour in different types of breast cancer.

3.2 Basal-Like vs. Non-Basal-Like Breast Cancer

3.2.1 Biological Background

In the current study we are comparing the gene expression profiles of basal-like and non-basal-like breast cancers with each other and normal breast cells. Basal-like cancers (BLC) account for 10

3.2.2 Materials and Methods

We are analysing the dataset provided by the NCBI library [54] which provides 7 samples for normal breast cells, 18 for basal-like and 20 for non-basal-like cancers. We ran 3 separate tests, comparing 2 breast cancer types in each (normal vs. basal-like, normal vs. non-basal-like, basal-like vs. non-basal-like). For that, we first separated the samples for each test and calculated expression averages for each type. We then removed probes with similar expression rates in each two sample groups in a test leaving us with 10000-16000 probes for each test instead of 54000 as in the original dataset.

We then ran the minimum spanning tree k-nearest neighbor algorithm on each test separately. Finally we analysed the results by focusing on the most differentially expressed genes in each test and the most central genes from the p53 network.

3.2.3 Results

The clustering results for normal vs. basal-like, normal vs. non-basal-like, and basal- vs. non-basal-like samples are displayed in figures 3.2, 3.9 and 3.10 correspondingly. We also located the most interesting genes from the p53 network and marked down the clusters in each of the tests (results in table 3.1). Not all genes were present in the dataset and not all of them were expressed differently enough to appear in our final results.

Normal breast cells vs. basal-like breast cancers

After calculating the ratios of average expression rates in the two cell types we identified the probes that displayed an approximately 3-fold difference. These gene names are listed in table 3.2.

We can see from the results that the most underexpressed probes in basal-like cancer cells compared to normal ones are all positioned in cluster 2. Taking a look at table 3.1

Central proteins(genes) from p53 network	Normal vs. basal cluster nr	Normal vs. nonbasal cluster nr	Basal vs. nonbasal cluster nr
p53 (TP53)	2	1	
Cdk1 (CDC2)	1,0	0	
Mdm2		4	
pRb (RB1)	2		3
PCNA	0	0	
DNA-PK (PRKDC)	1		
p21 (CDKN1A)	2		
E2F1-2-3	1,0,0	19,19,57	29,0,14
Abl (ABL1)	2		
Gadd45*		8	7
PARP*	1,2,19	1,4,10,40	3,6
ATM	2		
Cdc25A	1,9	8,19	0,1

Table 3.1: The most central proteins from the p53 protein-protein interaction network and their locations in test results.

Underexpressed in cancerous cells	Cluster nr	Overexpressed in cancerous cells	Cluster nr
hCG_25653	2	ART3	9
CITED1	2	CXorf61	0
FIGF	2	CENPA	0
SCUBE2	2	FOXM1	0
SCARA5	2		
AI492388	2		

Table 3.2: Probes with at least a 3 fold expression difference between normal breast cells and basal-like breast cancer cells.

we can see that several of the interesting p53 network genes are also located in cluster nr. 2: p53, pRb, p21, Abl, PARP and ATM. The fact that these genes are clustered together with the most differentially expressed probes further emphasizes the importance they have in the development of malignant tumors. Cdk1, PCNA, E2F1-2-3 and Cdc25A are clustered together with some of the most overexpressed probes in basal-like cancer cells.

Results from the centrality calculations for clusters 2 and 0 are displayed in table 3.3. The expression rates for some of the probes can be seen in figures 3.3 and 3.4. It is interesting to note by looking at the graphs and the centrality results that the genes that best distinguish between basal-like breast cancers and healthy cells are the most central ones in their clusters. TFAP2 is a transcription factor believed to stimulate cell proliferation and suppress terminal differentiation of specific cell types during embryonic develop-

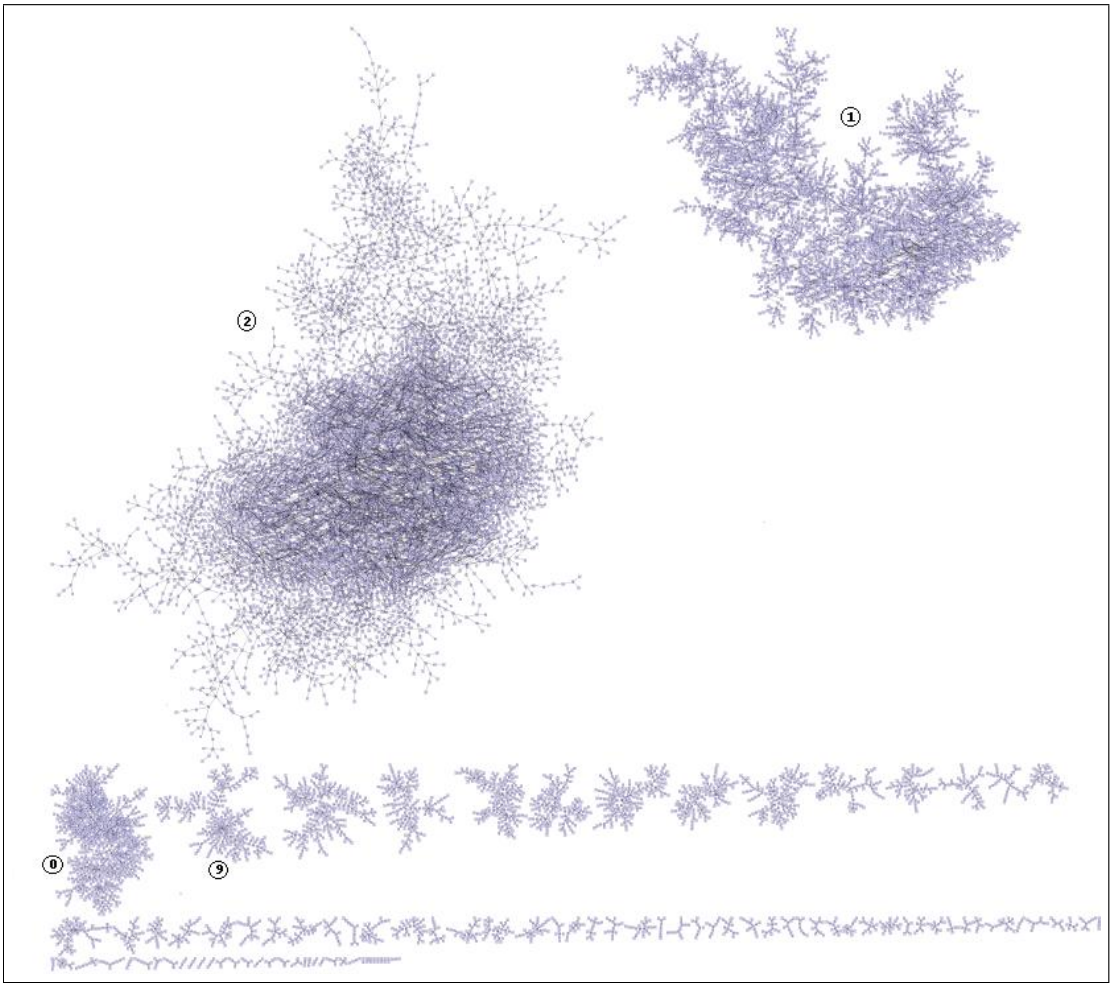


Figure 3.2: Clustering results for the gene expression analysis of normal breast cells and basal-like breast cancer cells

ment. It distinguishes perfectly between healthy and basal-like tumor cells, as well as other genes with short average path lengths like FIGF. TP53 (gene that encodes the p53 protein), which is among the least central in the cluster, behaves quite chaotically making it a bad discriminator for basal-like breast tumors. However, it is interesting to note that p53 levels are hardly ever similar to those of normal cells in any basal-like tumors. They are either too high or too low, but not normal.

Other great biomarkers for basal-like breast tumors include TOP2A and CDC2, which are some of the most central probes in cluster 0. TOP2A is an enzyme that functions as the target for several anticancer agents and a variety of mutations in this gene have been associated with the development of drug resistance. CDC2 (Cdk1) is familiar to us as one of the most central proteins in the p53-network. It is a highly conserved protein that is a key player in cell cycle regulation. Again we can make a conclusion that the probes that discriminate the best between normal breast cells and basal-like breast tumors can be found among the most central genes in the clusters with the most differentially expressed probes.

Cluster 2 results (8159 probes)			Cluster 0 results (777 probes)		
Position	APL	Probe name	Position	APL	Probe name
1	0.28345	TFAP2B	1	0.19292	TOP2A
2	0.28789	PNLIPRP3	2	0.20032	CDC20
3	0.28953	COL17A1	10	0.21914	FOXM1
5	0.29059	FIGF	11	0.21966	CDC2
81	0.32483	SCUBE2	12	0.22030	CENPA
86	0.32608	CITED1	59	0.26502	PCNA
148	0.33802	AI492388	62	0.26683	E2F3
188	0.34391	hCG_25653	416	0.43508	CXorf61
683	0.38600	SCARA5	555	0.49370	E2F2
2784	0.46242	ABL1			
3142	0.47260	PARP8			
3616	0.48587	PARP3			
4533	0.51016	ATM			
4881	0.52052	CDKN1A			
6738	0.58425	RB1			
7306	0.61669	PARP11			
7675	0.65077	TP53			

Table 3.3: Average path lengths of interesting probes

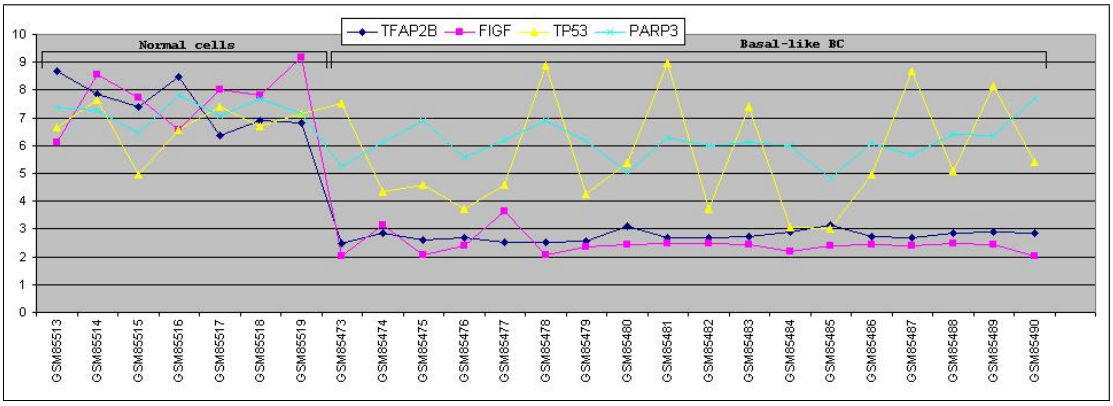


Figure 3.3: Expression rates in normal breast cells and basal-like breast tumors of the genes that are underexpressed in tumors

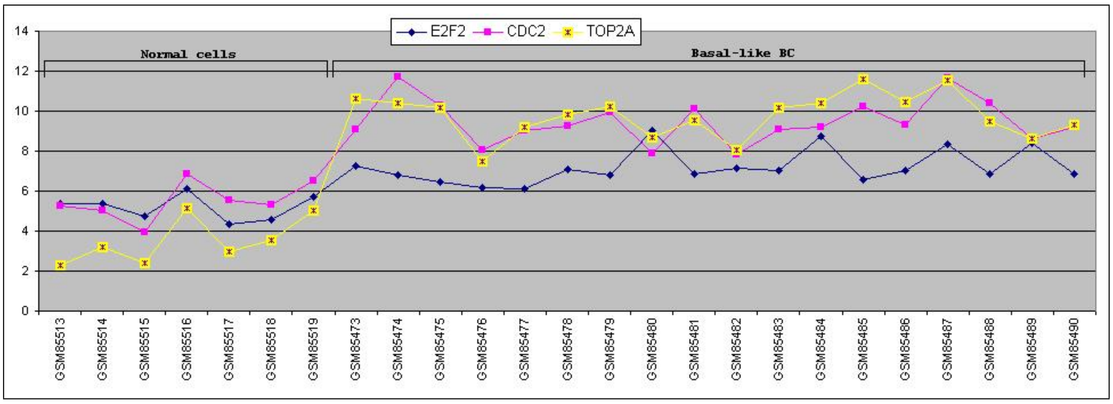


Figure 3.4: Expression rates in normal breast cells and basal-like breast tumors of the genes that are overexpressed in tumors

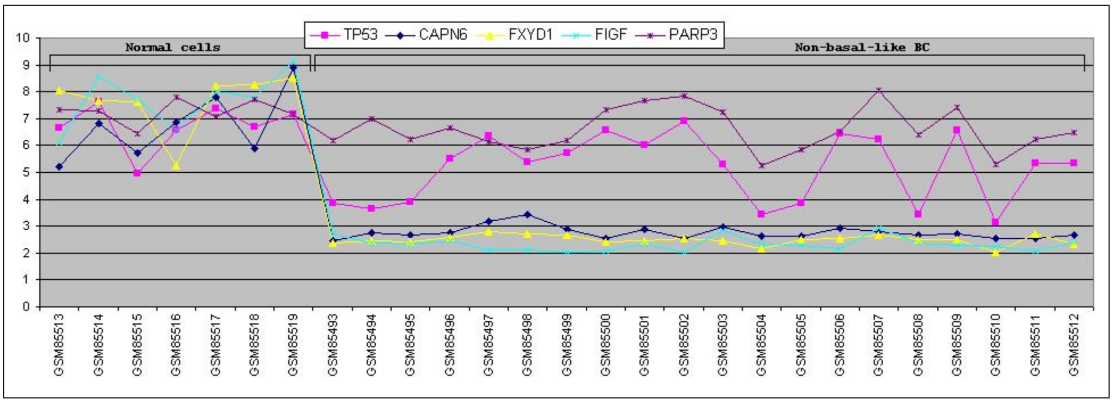


Figure 3.5: Expression rates in normal breast cells and non-basal-like breast tumors of the genes that are underexpressed in tumors

Normal breast cells vs. non-basal-like breast cancers

The most differentially expressed genes in non-basal-like breast tumors compared to normal cells are listed in table 3.4. According to this table the most underexpressed genes in cancerous cells seem to be in clusters 1 and 16, and the most overexpressed ones in clusters 0 and 27. Some of the most central genes from the p53-network appear also in cluster 1 (p53, PARP) and some in cluster 0(CDC2/Cdk1, PCNA), but none in clusters 16 and 27 and so we won't analyze these any further.

The results from the centrality calculator on clusters 0 and 1 are displayed in table 3.5 and figures 3.5 and 3.6. The best genes for differentiating between non-basal-like cancers and normal breast cells are CAPN6, FXYD1 and FIGF, which are all significantly underexpressed in tumors. CAPN6 is a calpain, a calcium-dependent cysteine protease

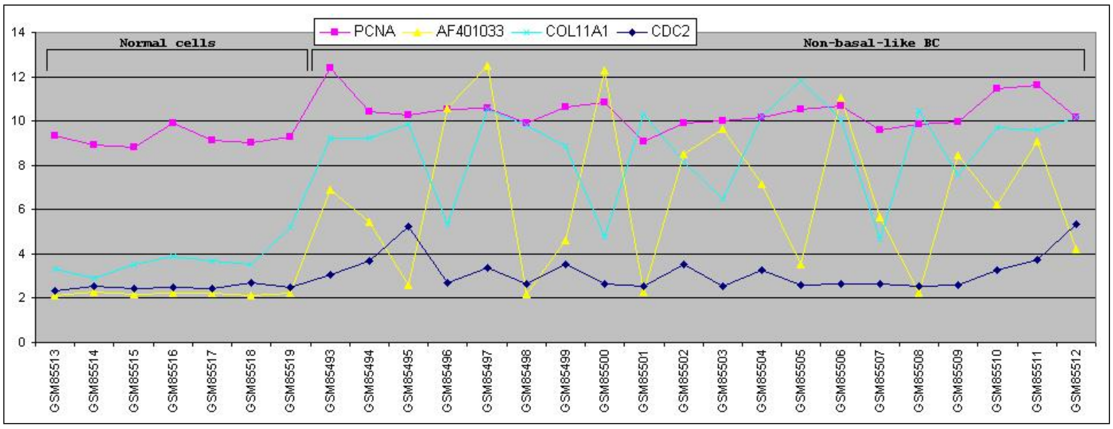


Figure 3.6: Expression rates in normal breast cells and non-basal-like breast tumors of the genes that are overexpressed in tumors

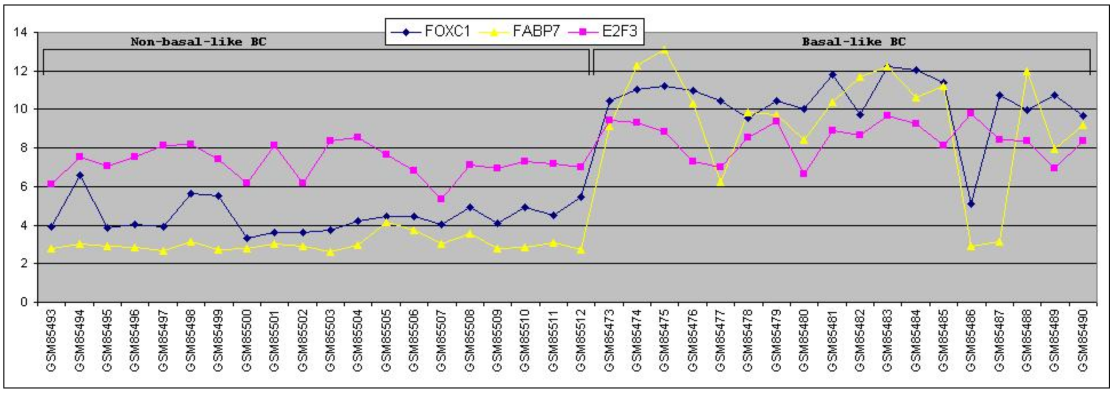


Figure 3.7: Expression rates in basal- and non-basal-like breast tumors of the genes that are overexpressed in basal-like cancers

Underexpressed in cancerous cells	Cluster nr	Overexpressed in cancerous cells	Cluster nr
SCARA5	1	AF401033	0
FIGF	1	COL11A1	0,27
FXYD1	1	A1376003	27
ROPN1	16		

Table 3.4: Probes with at least a 3 fold expression difference between normal breast cells and non-basal-like breast cancer cells.

involved in signal transduction in a variety of cellular processes. FXYD1 is thought to form an ion channel or regulate ion channel activity. FIGF is an endothelial growth factor.

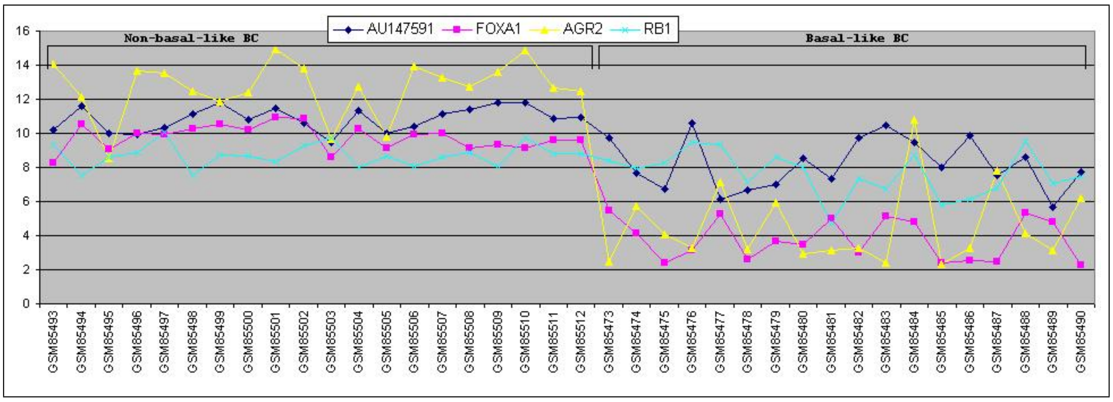


Figure 3.8: Expression rates in basal- and non-basal-like breast tumors of the genes that are underexpressed in basal-like cancers

Cluster 0 results (1270 probes)			Cluster 1 results (4187 probes)		
Position	APL	Probe name	Position	APL	Probe name
1	0.30644	CDC2	1	0.29420	CAPN6
40	0.37822	PCNA	3	0.29709	FXYD1
458	0.52773	AF401033	6	0.29958	FIGF
637	0.56233	COL11A1	221	0.38115	SCARA5
			964	0.46894	TP53
			1849	0.52536	PARP3

Table 3.5: Average path lengths of interesting probes in normal vs. non-basal-like cells

Yet again, the most central probes in cluster 1 are the best markers for distinguishing between normal cells and non-basal-like tumors. Cluster 0 however does not seem to be the best pick for separating normal cells from the tumors in consideration. The most central genes in the cluster, CDC2 and PCNA, are expressed quite similarly in the two groups and without major fluctuations. Even COL11A1 does not separate healthy cells from diseased in several samples, even though it is the best of the four.

Basal- vs. non-basal-like breast cancers

The probes that differentiate the best between basal- and non-basal-like breast cancers are listed in table 3.6. The most overexpressed probes in non-basal-like breast tumors compared to basal-like ones are all located in cluster 14, and underexpressed probes in clusters 3 and 26. Not many proteins from the p53-network were differentiating well between basal- and non-basal-like tumors, and only pRb, PARP and E2F are present in the clusters that we are interested in.

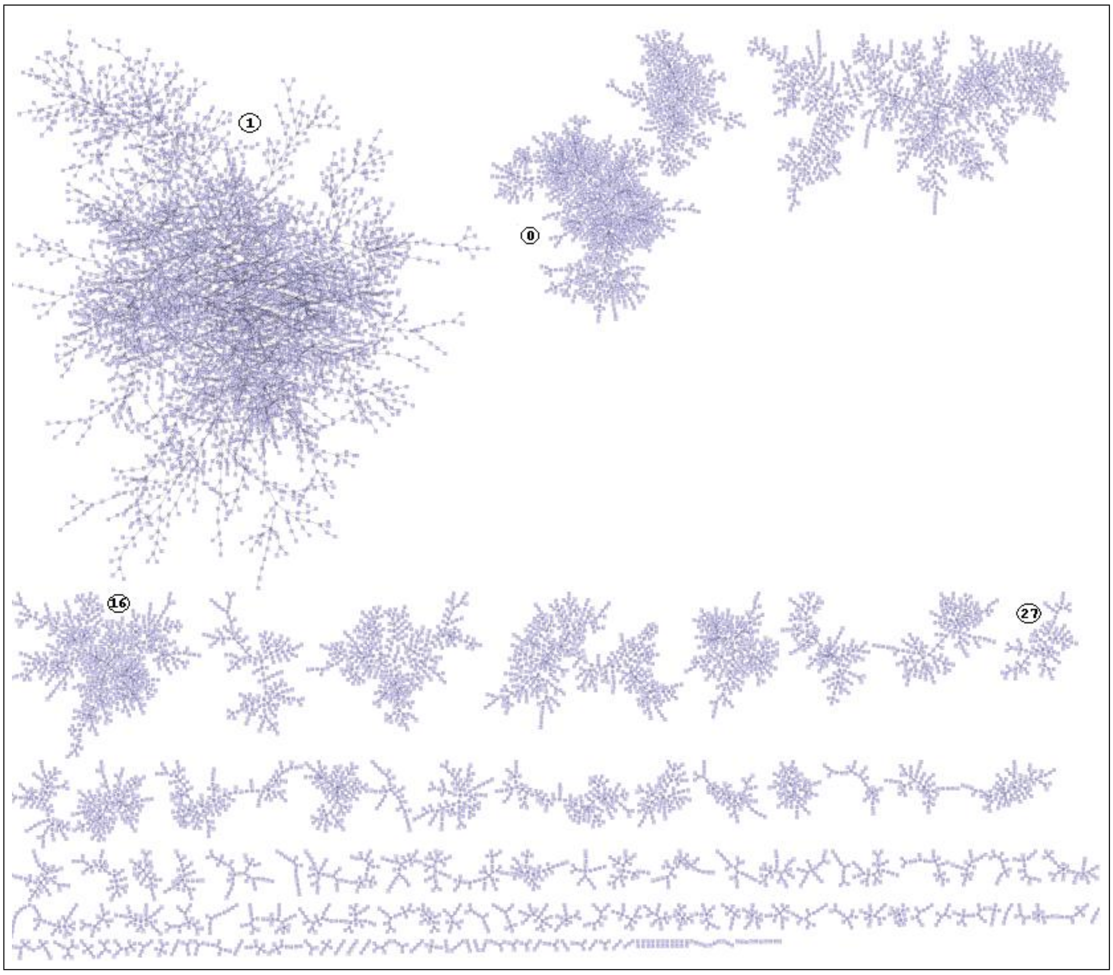


Figure 3.9: Clustering results for the gene expression analysis of normal breast cells and non-basal-like breast cancer cells

Underexpressed in non-basal cells	Cluster nr	Overexpressed in non-basal cells	Cluster nr
AF401033	26	ROPN1	14
AGR2	3	FABP7	14
FOXA1	3	ART3	14
AGR3	26	CXorf61	14
		SCRGI	14
		HORMAD1	14

Table 3.6: Probes with at least a 3 fold expression difference between basal- and non-basal-like breast cancer cells.

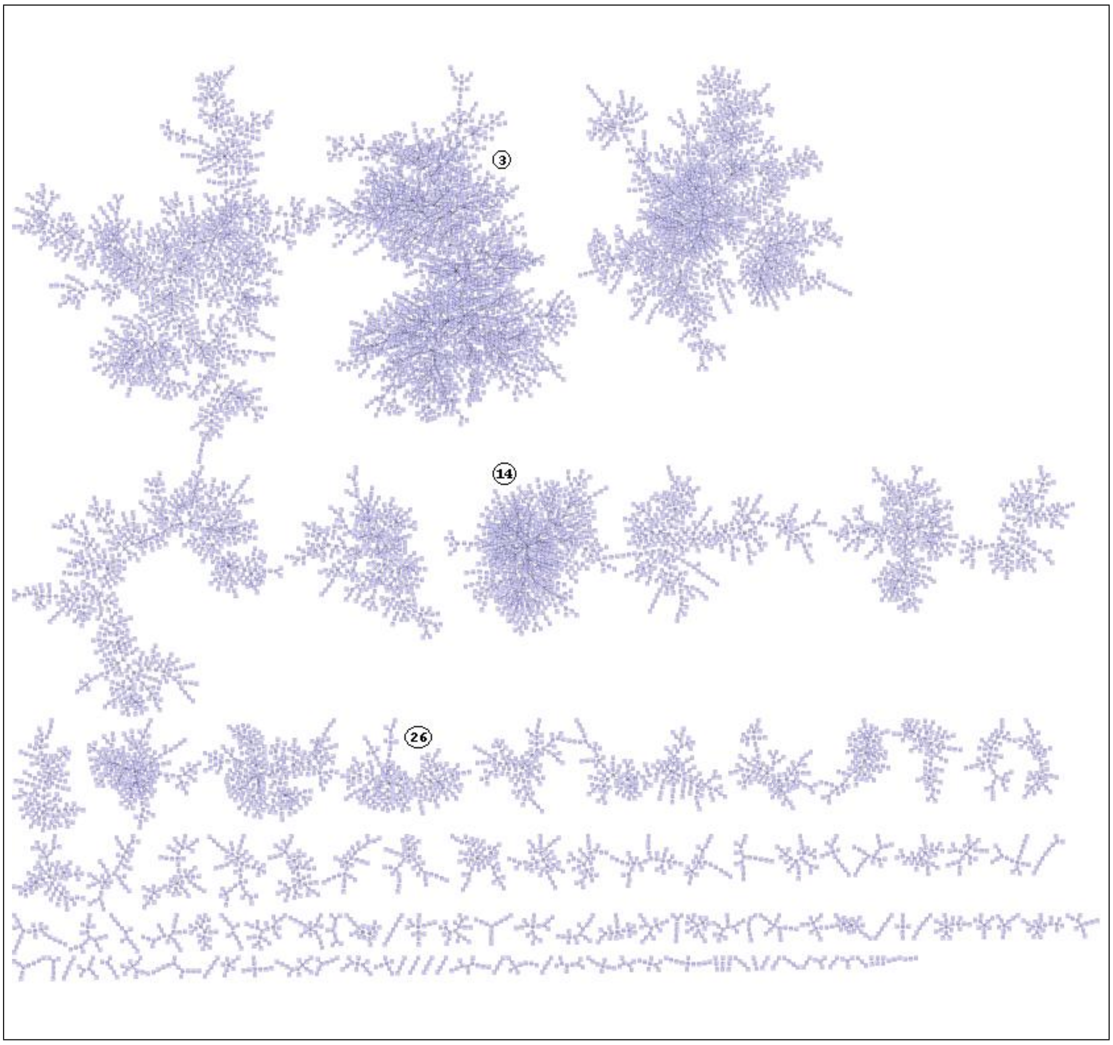


Figure 3.10: Clustering results for the gene expression analysis of basal- and non-basal-like breast cancer cells

When looking at the centrality calculation results in table 3.7 and the graphs with some of the probes in figures 3.8 and 3.7 we can yet again see that the most central genes in these clusters differentiate the best between different types of breast cancers. One of the best genes to differentiate between basal- and non-basal-like tumors is FOXA1 (Forkhead box protein A1). It is overexpressed in non-basal-like tumors (hormone induced breast cancers). In the past FOXA1 expression was always observed in human prostate carcinomas, which are mostly also hormone dependent.

Cluster 14 results (602 probes)			Cluster 3 results (1848 probes)		
Position	APL	Probe name	Position	APL	Probe name
1	0.41478	FOXC1	1	0.48252	AU147591
2	0.45433	FABP7	3	0.48871	FOXA1
12	0.51280	ART3	13	0.51409	AGR2
20	0.52288	SCRG1	901	0.70173	RB1
22	0.52657	ROPN1			
35	0.54140	CXorf61			
137	0.62578	HORMAD1			
144	0.62851	E2F3			

Table 3.7: Average path lengths of interesting probes (basal- vs. non-basal-like breast cancers)

3.2.4 Conclusions

It is visible from the results presented in this test that centrality can be used for analyzing gene expression data. We saw that the most central genes of certain clusters can be the best discriminators between different classes of data. It is also interesting to note how many of the p53 pathway genes came up among good distinguishers between normal breast cells and basal-like breast tumors. It might shed some light on why basal-like tumors are generally more resistant to popular chemotherapy drugs than hormone-dependent tumors.

3.3 Alternative Drugs for Breast Cancer Treatment

3.3.1 Motivation

Breast cancer is the second leading cause of cancer deaths in women today (after lung cancer) and is the most common cancer among women, excluding nonmelanoma skin cancers. According to the American Cancer Society, about 1.3 million women will be diagnosed with breast cancer annually worldwide and about 465,000 will die from the disease. One out of 8 women develop breast cancer during their lifetime.

Most cancer chemotherapy drugs come with the cost of severe side effects which range from temporary hair loss and nausea to longer term organ damage. The total economic cost of breast cancer in New South Wales, Australia in 2005 was estimated at 653600AUD per person, of which the financial cost is 64300AUD and the burden of disease is 589300AUD [42]. Drugs that would specifically target breast cancer cells could save patients from most side effects and hence also reduce the economic cost of cancer radically.

3.3.2 Biological Background

Breast cancers are often regulated by gonadal hormones like estrogen and/or progesterone. Therefore one possible treatment for such cancers is by blocking the hormone receptors on affected cells. People with hormone positive tumors have better chances for survival and their treatment is also less agressive than that for patients with hormone negative tumors [48].

3.3.3 Materials and Methods

The US National Cancer Institute (NCI) 60 human tumour cell line anticancer drug screen (NCI60) was developed in the late 1980s as an in vitro drug-discovery tool intended to supplant the use of transplantable animal tumours in anticancer drug screening [38]. It holds a large amount of data describing the cell growth inhibition effects of specific drugs on different cancer cell lines.

In our research we focused on the 7 breast cancer cell lines and used the other cancer cell lines as controls, as it has not been possible to create immortalised, i.e. indefinitely proliferating cell lines from normal cells. The objective of this work was to find a drug or a drug combination that might be a good alternative for current popular breast cancer chemotherapy drugs.

The effectiveness of cancer treatment depends on many factors: age, lifestyle and perhaps most importantly, the genetic signature of cancer cells. That explains why most medications generally work for only up to 15-20% of patients. In addition, most popular treatments come with severe side effects due to the inability of the medications to differentiate between normal and cancerous cells.

The graph in figure 3.11 shows cell growth inhibition properties for some of the most popular breast cancer treatment drugs. It is based on the GI50 data from the National Cancer Institute Developmental Therapeutics program. The values are negative logarithms of drug concentrations that inhibit the growth of the specific cell line by 50% [55]. In short, the bigger the number the better the drug is for controlling the growth of specific cells.

It is visible from the graph above that most drugs work more or less similarly for all cancer cell lines. In fact, most cancer treatment drugs used today cannot differentiate between cancerous and healthy cells, but instead target all fast dividing cells, including hair growth and white blood cells. So not only does hair fall out and does not grow back

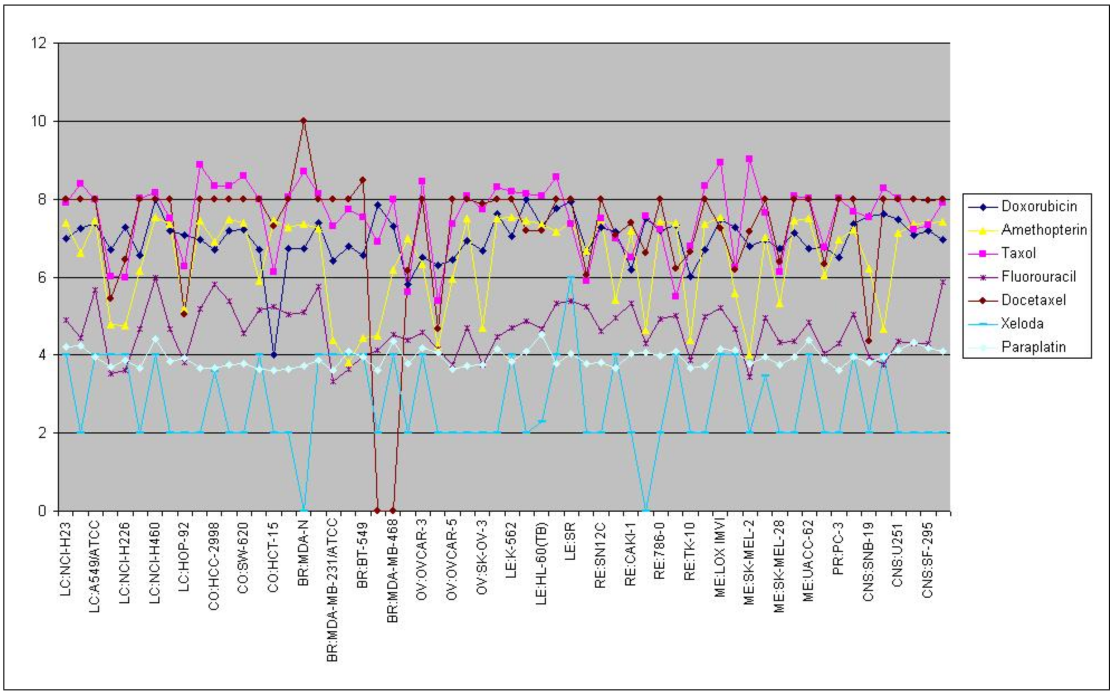


Figure 3.11: Inhibition properties of popular breast cancer chemotherapy drugs on different cancer cell lines

until after treatment has stopped, but also the patients' immune system is diminished leaving them more susceptible to colds, flu and other infectious diseases.

Another problem with chemotherapy is resistance to treatment. The mechanisms causing resistance to chemotherapeutic drugs in cancer patients are poorly understood. Certain mutations in essential proteins like the tumor suppressor p53 can often be blamed. In case of hormone induced cancers like breast and prostate cancer resistance develops over time.

Even though breast and prostate cancers arise in anatomically different organs, both organs need sex steroids for their development. Therefore the tumors that develop from them are often hormone-dependent and biologically similar. The main hormones that these cancers depend on are oestrogen and androgen. A common trait of hormone-dependent tumors is that even though they initially respond to hormone therapy they eventually develop resistance to the drugs [35]. Therefore it is of great importance to find combinations of drugs to overcome this problem.

Our goal for this project is to find alternatives for the common chemotherapy drugs for breast cancer. We are looking for substances that are less damaging for healthy cells while still effectively attacking cancerous cells. That would in theory reduce bad side effects which can often be so severe, that patients have to stop treatment altogether.

In order to analyze the NCI60 dataset with our clustering tools we had to make certain modifications to the data. We removed all drugs where standard deviation was below 0.01, i.e. the drugs that worked with the same efficiency for all cancer cell lines. We also removed all drugs and cell lines where at least half of the values were missing. We did not normalize the data, but we replaced missing values with drug response averages over all cell lines in order for the programs to work correctly.

We added a dummy row to the dataset representing a perfect drug for treating breast cancer. We set the effectiveness of the drug to the maximum for all breast cancer cell lines and minimum for the rest. The intention of this was for the clustering algorithm to find the best compounds for treating breast cancer and save them as the nearest neighbors to the dummy row.

3.3.4 Description of relevant cancer cell lines

In this study we are focusing on breast and prostate cancer cell lines. In order to understand the results better I provide here a short description about each of these cell lines.

MDA-N MDA-N was derived from MDA-MB435. By gene expression pattern, MDA-N and MDA-MB435 are very similar.

MCF7 Estrogen receptor positive cell line, breast carcinoma.

HS 578T Highly metastatic, growth inhibited by retinoids (compounds related to vitamin A); infiltrating ductal carcinoma (a type of tumor that primarily presents in the ducts of a gland).

BT-549 Infiltrating ductal carcinoma.

T-47D Estrogen-dependent, estrogen receptor positive. Infiltrating ductal carcinoma.

MDA-MB-231/ATCC Estrogen-independent, estrogen receptor negative cell line; adenocarcinoma (cancer of epithelia originating in glandular tissue)

MDA-MB-468 Estrogen receptor negative cell line; adenocarcinoma.

NCI/ADR-RES This cell line was initially believed to be derived from a breast tumor cell line MCF-7, but it was later found to share a large number of karyotypic abnormalities with the ovarian tumor cell line OVCAR-8, that it is now believed to be derived from [61].

MDA-MB-435 This cell line was derived at M.D. Anderson in 1976 from a 31-year old woman with a history of breast cancer. Recently it has been shown however that the MDA-MB-435 and M14 (melanoma) cell lines are essentially identical with respect to cytogenetic characteristics as well as gene expression patterns [32]. It has not been determined, however, whether the melanoma-like properties of the MDA-MB-435 cell line are the result of misclassification or due to transdifferentiation (a non-stem cell transforms into a different type of cell) to a melanoma-like phenotype.

PC-3 Originally derived from advanced androgen independent metastasized prostate cancer (bone metastasis). PC3 cells have high metastatic potential compared to DU145 cells which have a moderate metastatic potential.

DU-145 It is a prostate cancer cell line derived from brain metastasis. It is not hormone sensitive.

3.3.5 Results

As a first step we ran the clustering algorithm on the dataset using three different distance metrics: Eucleidean, Pearson and Spearman correlation metrics. We got different scale-free networks for each metric. The results are displayed in figures 3.12, 3.13 and 3.14.

It is visible from the results that all three metrics produce rather different clustering results. The next step is to take a closer look at the compounds that have been identified by the clustering algorithm as the ones with the most similar behaviour to that of our ideal dummy drug. We generated graphs 3.15(a), 3.15(b) and 3.15(c) displaying the response rates of the closest compounds of our dummy drug according to each distance metric.

Both Euclidean and Pearson distance metrics identified enhydrin and kedarcidin as the best compounds for curing breast cancer while leaving other cancer cell lines relatively untouched. Methanesulfinic acid and hydramycin were identified as the third best choices, but one look at the charts reveals that both of them work uniformly similarly for all cell lines.

The two best results when using the Spearman correlation coefficient are compounds that we could not find any information about. But as both of them work uniformly similarly for all cell lines we are not studying them any further.

The behaviour of enhydrin and kedarcidin on all cancer cell lines can be seen in 3.16(b). For comparison we also added a graph showing the effects of two common breast cancer drugs, docetaxel and imatinib in 3.16(a). As can be seen from these figures, docetaxel and imatinib target quite uniformly all cancer cell lines. It can therefore be assumed that they have a similar effect on other rapidly growing cells, resulting in

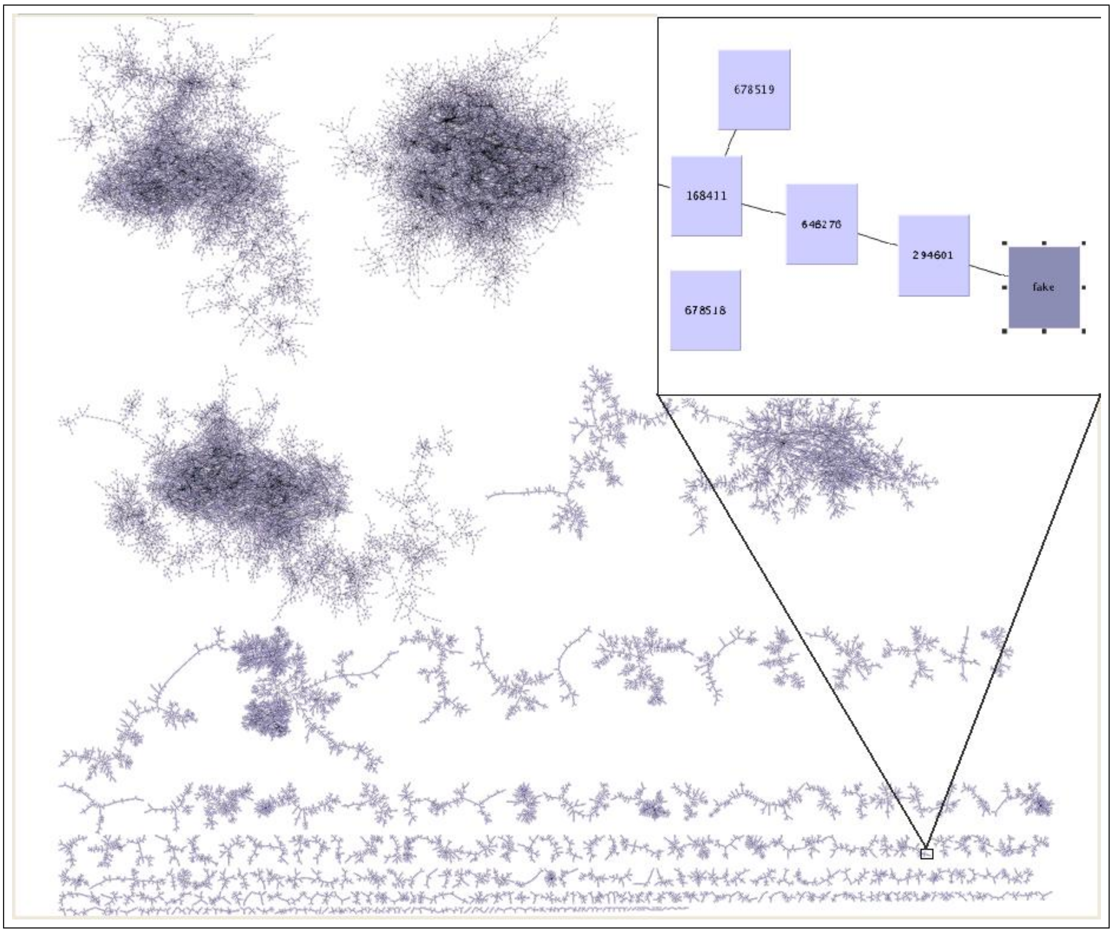


Figure 3.12: Clustering results using Euclidean distance as the metric

severe side effects in patients.

Taking a look at enhydrin and kedarcidin, we notice that they effectively inhibit the growth of all breast and prostate cancer cell lines that we have information about, and a single melanoma and ovarian cancer cell lines. The specificity of how enhydrin and kedarcidin target cancer cells suggests the possibility that other fast growing healthy cells would be left relatively untouched. That would result in less severe side effects of chemotherapy.

An interesting fact about the melanoma cell line MDA-MB-435 and the ovarian cell line ADR-RES is that they were mistaken for breast cancer cell lines and only recently after gene expression studies classified as melanoma and ovarian cancer cell lines correspondingly ([43], [44]). The study by Rae et al. [32] supports the hypothesis that while MDA-MB-435 cells may originally have been of breast cancer origin, sometime after their establishment the cells were either contaminated with M14 melanoma cells,

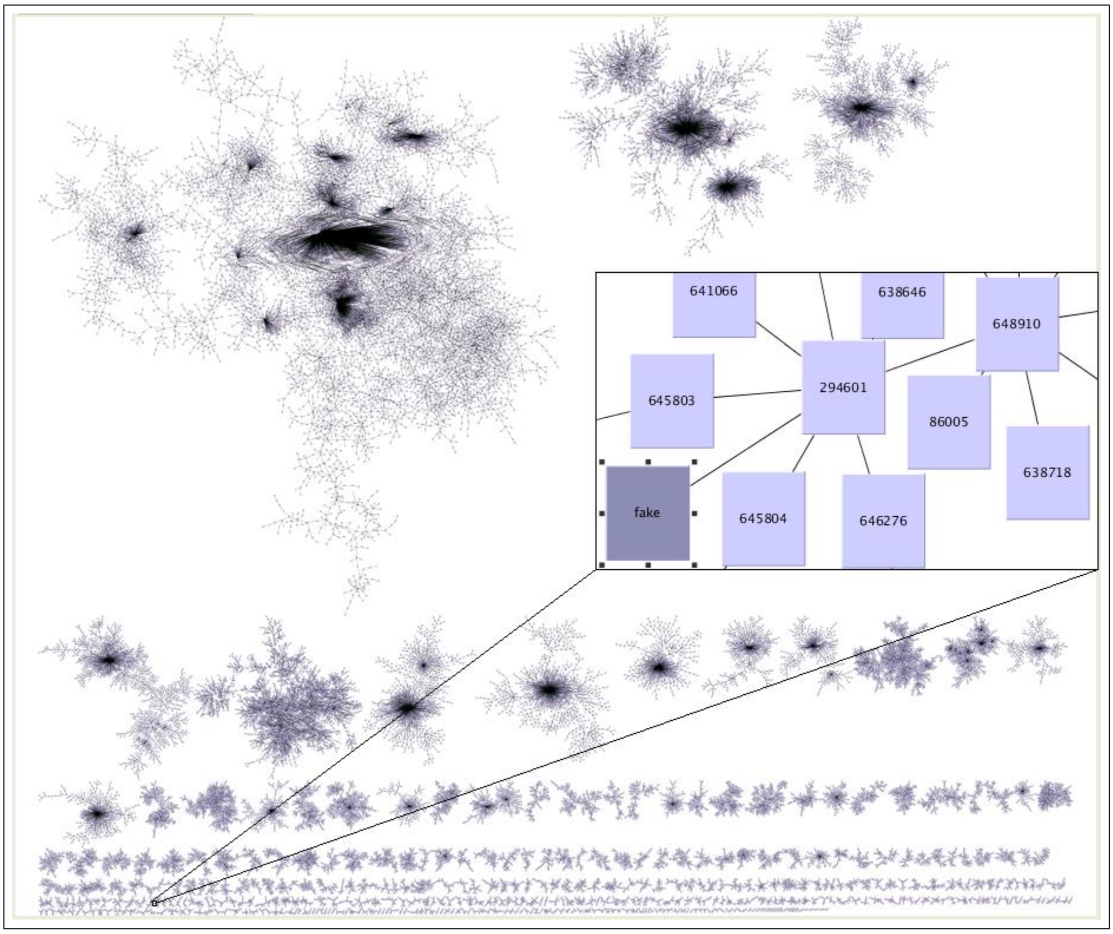


Figure 3.13: Clustering results using Pearson correlation coefficient as the metric

which subsequently overgrew and replaced the breast cells, or some labelling error or other accident led to the misidentification of the culture. Considering that enhydrin and kedarcidin do not show much effect on the M14 melanoma cell line whereas displaying strong inhibition ability on MDA-MB-435, the results of our study suggest that MDA-MB-435 still has some similarity with breast cancer cell lines, and therefore might be of breast cancer origin.

It is tempting to explain the behaviour of enhydrin and kedarcidin by referring to the fact that breast and prostate cancers are mostly sex hormone dependent. It is therefore a logical assumption that enhydrin and kedarcidin somehow regulate hormone binding or inhibit hormone production. However, taking a look at the cell line descriptions it is obviously not as simple as that. Most breast and prostate cancer cell lines in NCI-60 database are not hormone-dependent.

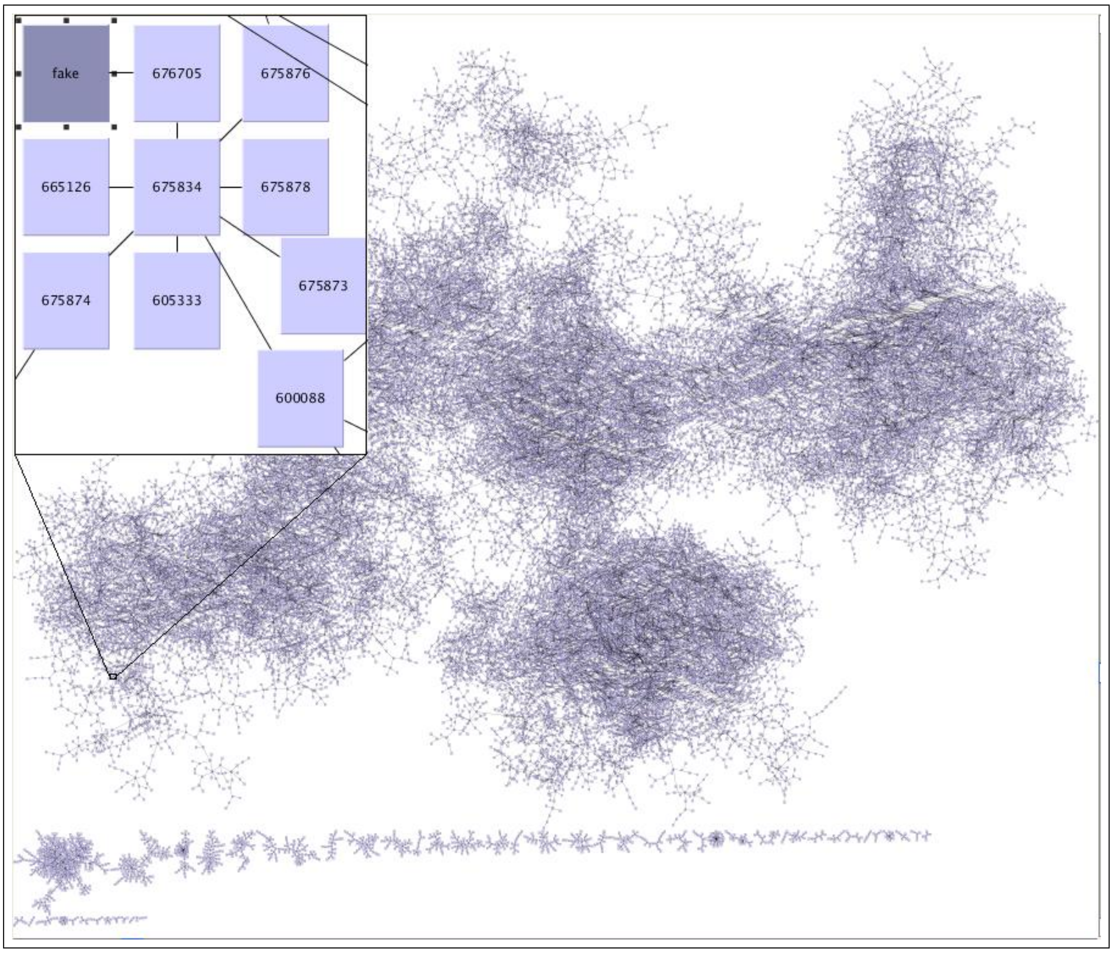
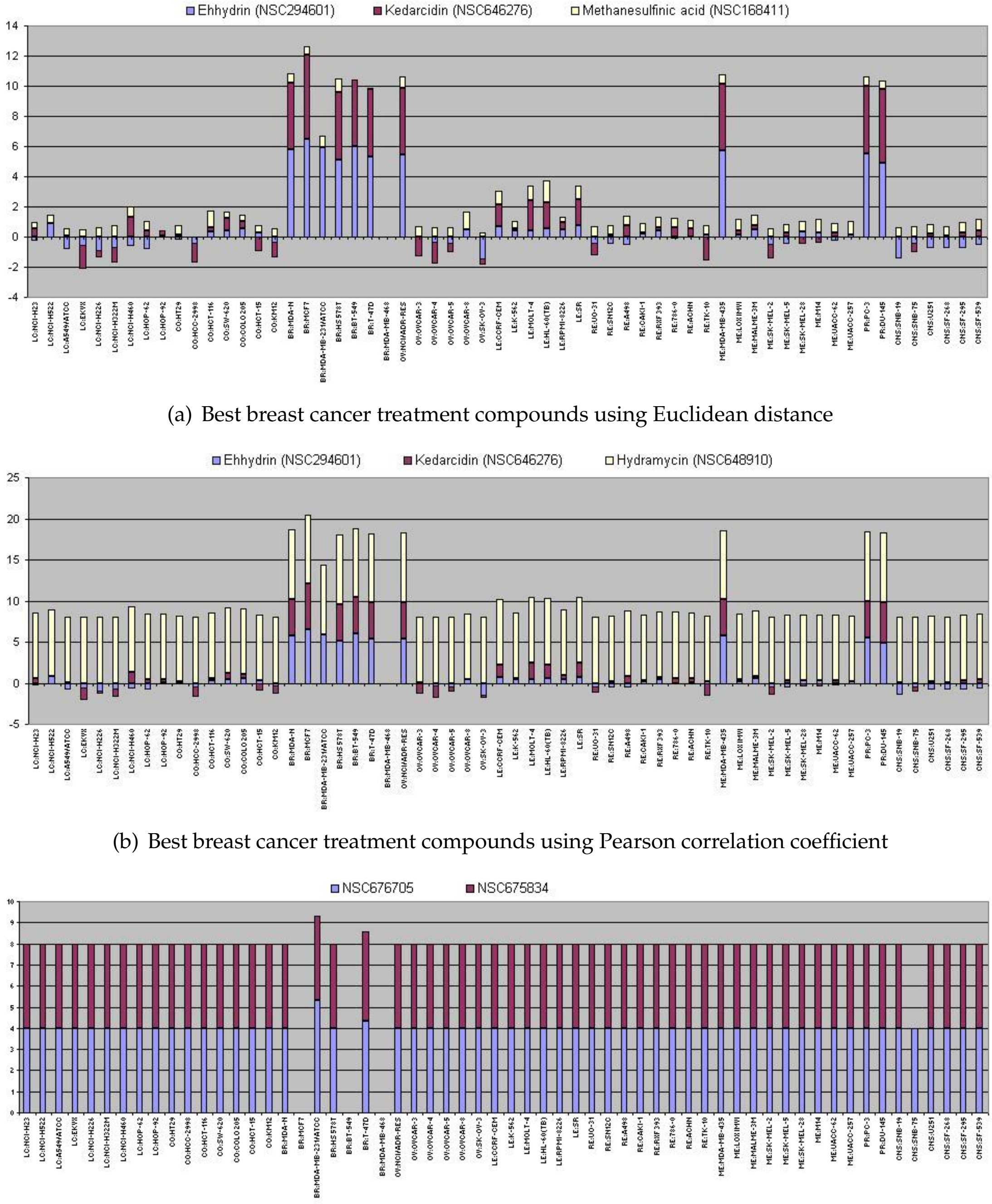


Figure 3.14: Clustering results using Spearman correlation coefficient as the metric

Centrality

Centrality studies for drug responses might at first glance seem useless and puzzling. Why would we be interested in finding the drugs that behave in the most average manner? Thinking about the meaning of clusters in the context of drug effects on different tumors provides us with an answer.

Clustering helped us retrieve all compounds that have similar cancer cell growth inhibition properties as our dummy drug. Given that cancer is genetically highly variable it is fairly safe to assume that a single drug will not be effective for curing different tumor types, even if they are all classified as breast cancers. Therefore therapies consisting of a combination of drugs are an attractive proposition. In addition to using compounds that are working well on test samples it makes sense to include drugs that work more generally with the hope that they work better for outliers. Looking at the

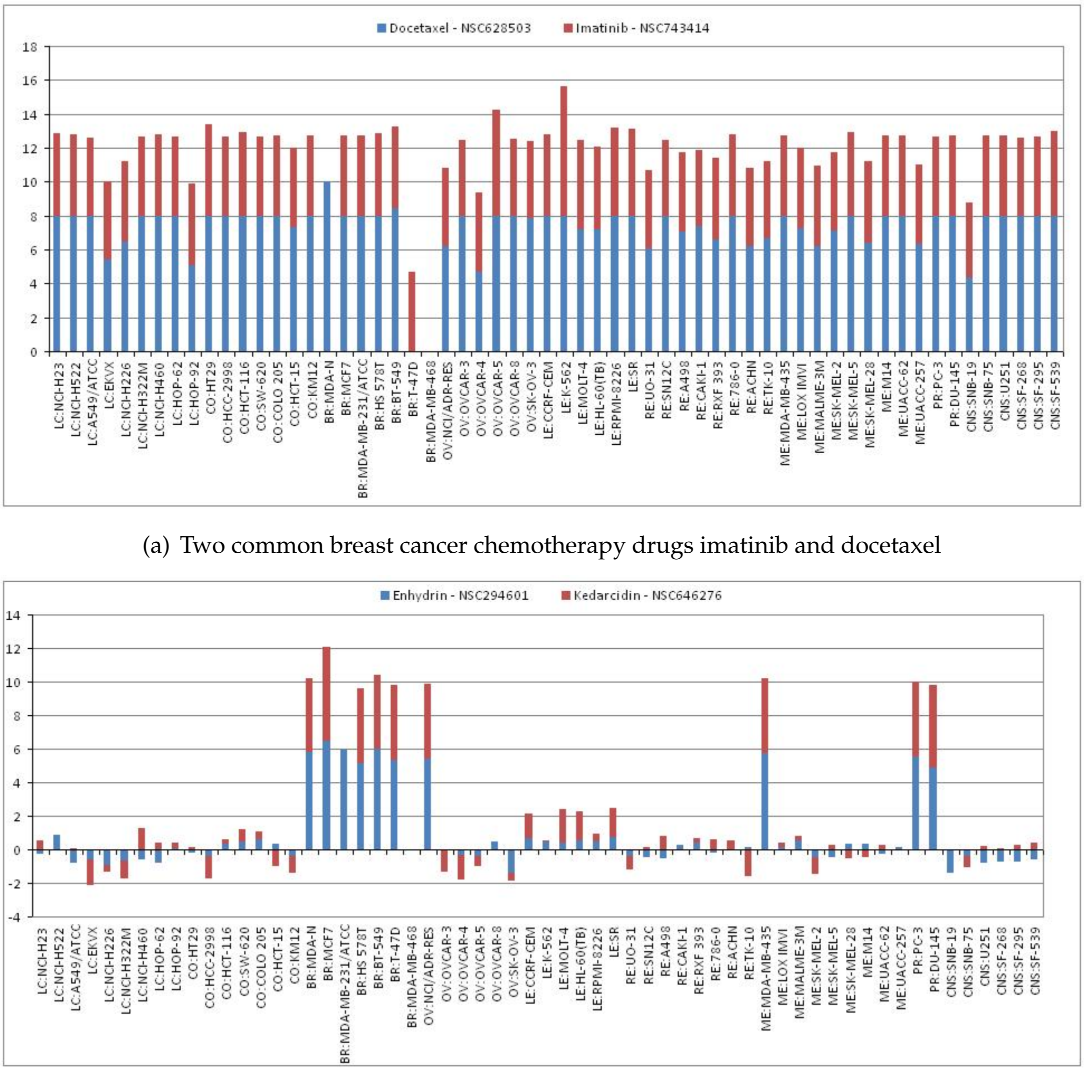


(a) Best breast cancer treatment compounds using Euclidean distance

(b) Best breast cancer treatment compounds using Pearson correlation coefficient

(c) Best breast cancer treatment compounds using Spearman correlation coefficient

Figure 3.15: Best breast cancer treatment compounds



(a) Two common breast cancer chemotherapy drugs imatinib and docetaxel

(b) Our proposed compounds enhydrin and kedarcidin

Figure 3.16: Comparision of in vitro cell growth inhibition properties for common breast cancer chemotherapy drugs and our proposed compounds over all cancer cell lines

most central compounds in the cluster is a good start as in a sense they generalize the clusters.

As Spearman correlation did not prove to be a useful clustering metric for this problem, we are only focusing on Euclidean distance and Pearson correlation for centrality studies.

Using Euclidean distance as the clustering metric our fake drug was clustered together with 71 other compounds. The output from the centrality calculator is given in table 3.8.

Centrality index	APL	NSC number	Compound Name
1	4.09542	636194	NOR 5H10
2	4.09985	636208	FRI 3C3
3	4.24627	636203	FRI A12
60	12.57550	294601	Enhydrin
66	18.21825	646276	Kedarcidin
72	64.36332	fake	

Table 3.8: Average path lengths (APL) for clustering results using Euclidean distance as a metric

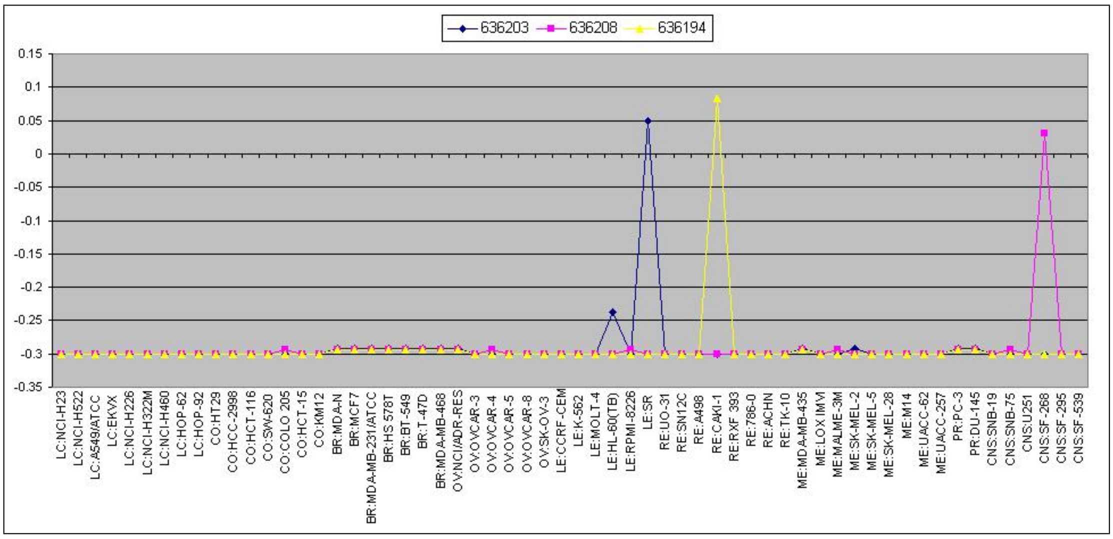


Figure 3.17: Three most central compounds using Euclidean distance as a clustering metric

The graph displaying the inhibition properties for the top 3 most central compounds can be seen in figure 3.17. When looking very carefully we can observe a slight bump for all breast and prostate cancer cell lines, and also specific melanoma lines. As the inhibition properties for all these substances are too small to be considered interesting we will not look into them any further. Our fake drug is the least central compound in this cluster, and enhydrin and kedarcidin are also among the last which emphasizes the uniqueness of these drugs.

Centrality calculations for the clustering results using Pearson correlation as a clustering metric can be found in table 3.9. This time the cluster consists of 25 different compounds including our fake drug.

Centrality index	APL	NSC number	Compound Name
1	0.45192	294601	Enhydrin
2	0.46287	86005	Nogalamycin compound A
3	0.54713	646276	Kedarcidin
25	2.12858	fake	

Table 3.9: Average path lengths (APL) for clustering results using Pearson correlation as a metric

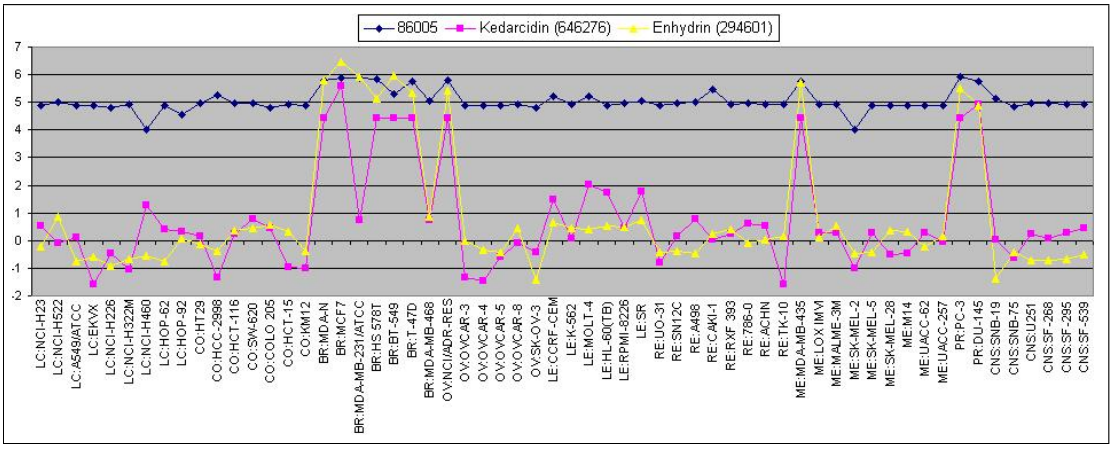


Figure 3.18: Three most central compounds using Pearson distance as a clustering metric

The results using Pearson correlation coefficient as a clustering distance metric differ considerably from the previous results using Euclidean distance as a metric. Enhydrin is the most central compound in this cluster, nogalamycin compound A following closely behind and kedarcidin coming third. Our fake drug is again positioned as the least central substance. The graph displaying the inhibition properties for the 3 most central compounds can be seen in figure 3.18. Enhydrin and kedarcidin are both very effective for inhibiting breast and prostate cell lines from growing whereas having only a little effect on other cell lines. Nogalamycin compound A is cytotoxic to all cancer cell lines, being even more effective in case of breast and prostate cancers. When added in small doses to enhydrin or kedarcidin it might help beat cancers that these two drugs on their own cannot. All these three compounds are covered in more detail below.

Enhydrin

Enhydrin is a sesquiterpene lactone which is found in two plants native to the southeastern United States, *Magnolia grandiflora* (Southern Magnolia) and *Smallanthus uvedalius* (known as Bear's foot). Enhydrin is also a major component of *Smallanthus sonchifolius* (popularly known as yacón) leaf extracts. Yacón is a perennial herb endemic to the eastern Andes of South America (from Venezuela to Argentina), but it has also been successfully cultivated in Italy, France, Germany, USA, Czeck Republic, Russia and Japan.

Ethnomedicinal records from Native American Cherokee note the use of Bear's Foot for its analgesic properties and Southern Magnolia for the treatment of fever, diarrhea, rheuma and arthritis. Feltenstein et. al. have shown that enhydrin might indeed be useful in the treatment of inflammation and pain [10]. In addition to that, enhydrin shows anti-fungal and antimicrobial activities [18], and is an anti-diabetic agent and an important component of pharmaceutical formulations [19]. We did not find any references to a study conducted with enhydrin in relation to breast cancer though.

Nogalamycin compound A

Nogalamycin compound A is an antracycline antibiotic used in cancer chemotherapy. It is derived from Streptomyces bacteria. These compounds are used to treat a wide range of cancers, including leukemias, lymphomas, and breast, uterine, ovarian, and lung cancers. The anthracyclines are some of the most effective anticancer treatments ever developed and are effective against more types of cancer than any other class of chemotherapy agents. Their main adverse effects are heart damage (cardiotoxicity), which considerably limits their usefulness, and vomiting [45].

Kedarcidin

Kedarcidin is a potent antitumor antibiotic chromoprotein, composed of an enediyne-containing chromophore embedded in a highly acidic single chain polypeptide [41]. It has attracted several research teams since 1990 and it has been generally regarded as having potential action of leukaemia and melanoma. Our findings seem to correlate well with this view (there are clear signs of differential inhibition behaviour of some Leukaemia cell lines (LE:CCRF-CEM, LE:MOLT-4, LE:HL-60, LE:RPMI-8226, LE:SR) and in one melanoma (ME:MDA-MB-435).

3.3.6 Conclusions

The exploration of mass datasets such as the NCI60 dataset provides a useful avenue for exploration of possible agents without incurring the temporal and financial costs of a similar wet-lab exploration. We have discovered two possible drugs for treating breast and prostate cancer that have given extremely promising results from *in vitro* studies. It is of great importance to continue studying the effects of enhydrin and kedarcidin on human body. As enhydrin has been used for centuries as a major constituent of yacón leaves in herbal medicine it has proved to be nontoxic, at least in small amounts. If it worked against breast and prostate cancer it would result in considerably less averse side effects for treating these cancers than chemotherapy drugs generally come with. Also, it would be interesting to find out why enhydrin and kedarcidin target these specific cell lines and not others. To some extent it could be explained by the cancers' dependence on sex hormones like oestrogen and androgen, but not all of these cell lines are hormone-induced. Actually, only 2 breast and 1 prostate cancer cell lines are hormone dependent and there are equally 2 breast and 1 prostate cancer cell lines that are hormone independent.

4 Applications

There are several applications that have been used throughout this thesis. I will describe here shortly some that I have written while working on this project.

4.1 The APLProject

This application was written for calculating average path lengths for all nodes of a graph as was described in chapter two. The opening page of the program is shown below (Figure 4.1). Accepted input file formats for this program are .vna and .net. The .net file format is in more detail described in [53]. The .vna file format simply describes all links in the graph by giving the origin and destination vertex names, states whether a link between them exists and how strong the connection is (Table 4.1).

from	to	exists	strength
p53	CycA	1	4
p53	RPA	1	1

Table 4.1: An example of a vna file

The opened file is displayed in the edit box and the combo box with the choice of algorithms becomes enabled (4.2). The output from Dijkstra and Floyd-Warshall is the same, the only difference is the running time for large graphs. Dijkstra in general performs slightly faster.

When ”Calculate average path lengths” is pressed, the output from the program is displayed (4.3). On top there is the distance matrix that was produced from the input graph, and further below the vertex names with their average path lengths are displayed, ordered according to centrality. Both the matrix and the centrality list can be saved separately by selecting one of the buttons above, or the whole output can be saved by choosing “File → Save” or “File → Save As”.

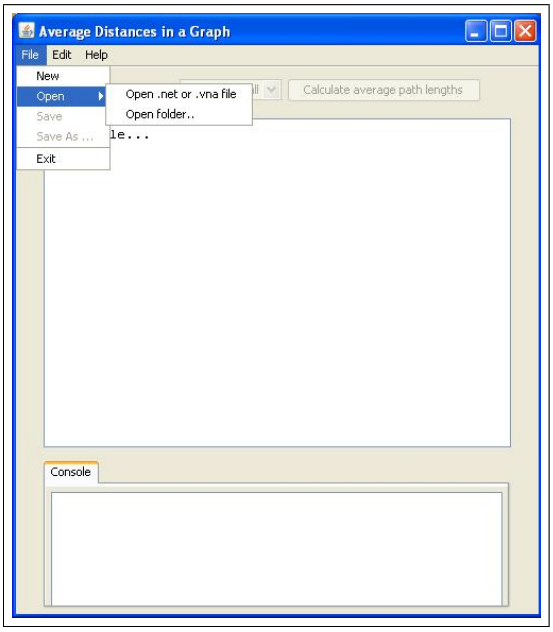
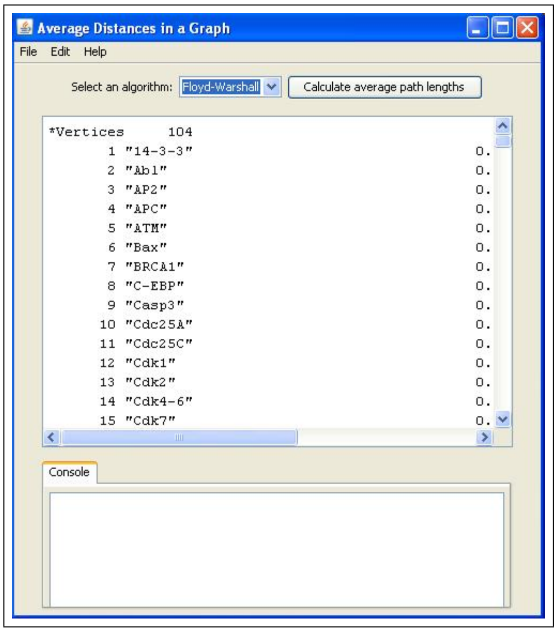


Figure 4.1: The APL project opening page

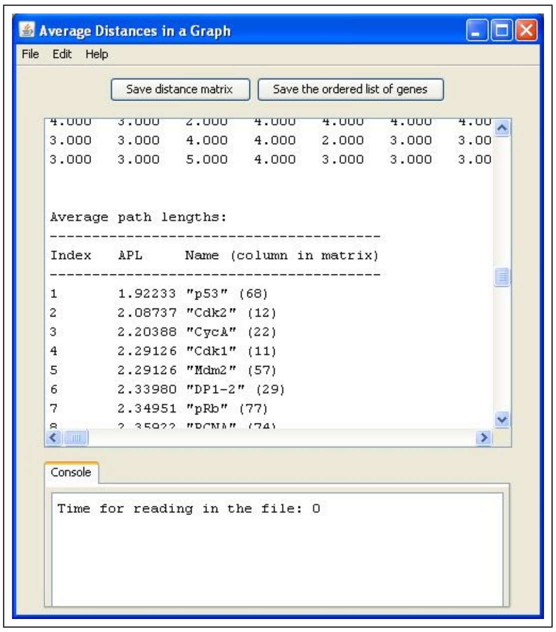
4.2 GeneFinder

GeneFinder is a simple tool for finding a list of genes from a folder or file and printing out all the rows where the genes occurred. It resembles the grep application, but has a graphical user interface.


Average Distances in a Graph
File Edit Help
Select an algorithm: Floyd-Warshall
Calculate average path lengths
Console

```
*Vertices     104
       1 "14-3-3"                             0.
       2 "Abl"                                0.
       3 "AP2"                                0.
       4 "APC"                                0.
       5 "ATM"                                0.
       6 "Bax"                                0.
       7 "BRCA1"                              0.
       8 "C-EBP"                              0.
       9 "Casp3"                              0.
      10 "Cdc25A"                             0.
      11 "Cdc25C"                             0.
      12 "Cdk1"                               0.
      13 "Cdk2"                               0.
      14 "Cdk4-6"                             0.
      15 "Cdk7"                               0.
```

Figure 4.2: The APLProject

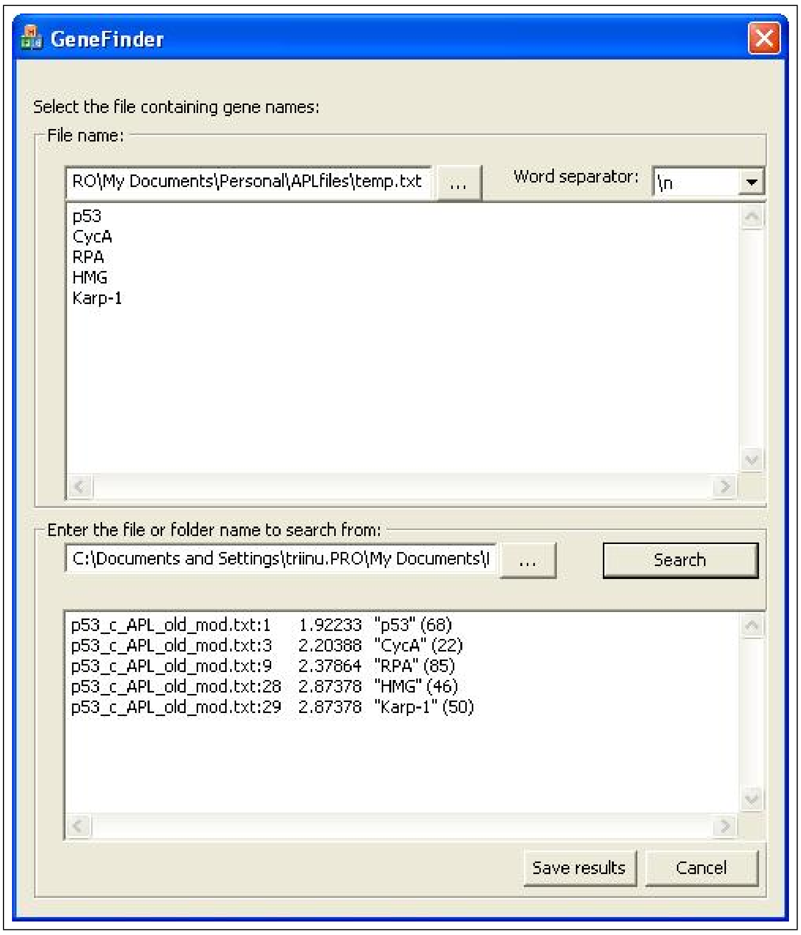


```
4.000   3.000   2.000   4.000   4.000   4.000   4.00
3.000   3.000   4.000   4.000   2.000   3.000   3.00
3.000   3.000   5.000   4.000   3.000   3.000   3.00


Average path lengths:
----------------------------------------
Index   APL     Name (column in matrix)
----------------------------------------
1       1.92233 "p53" (68)
2       2.08737 "Cdk2" (12)
3       2.20388 "CycA" (22)
4       2.29126 "Cdk1" (11)
5       2.29126 "Mdm2" (57)
6       2.33980 "DP1-2" (29)
7       2.34951 "pRb" (77)
8       2.35922 "PCNA" (74)
```

```
Time for reading in the file: 0
```

Figure 4.3: The APLProject



Select the file containing gene names:

File name: RO\My Documents\Personal\APLfiles\temp.txt

Word separator: \n

```
p53
CycA
RPA
HMG
Karp-1
```

Enter the file or folder name to search from: C:\Documents and Settings\triinu.PRO\My Documents\l

```
p53_c_APL_old_mod.txt:1    1.92233  "p53" (68)
p53_c_APL_old_mod.txt:3    2.20388  "CycA" (22)
p53_c_APL_old_mod.txt:9    2.37864  "RPA" (85)
p53_c_APL_old_mod.txt:28   2.87378  "HMG" (46)
p53_c_APL_old_mod.txt:29   2.87378  "Karp-1" (50)
```

Figure 4.4: GeneFinder

5 Conclusion

5.1 Future directions

There are several things left to do regarding the current project. It is essential to continue studying enhydrin and kedarcidin, and perhaps also hydramycin as possible drugs for treating breast cancer. So far we only have *in vitro* results about their cell growth inhibition properties which are hardly reliable as proof for curing breast cancer in humans. Several *in vivo* experiments need to be conducted in order to get a better idea of the effects these compounds might have on human body. It takes a lot of time and money to run such tests, but the ever increasing cost of treating cancers should be a good enough inducement.

Another possible direction for future work is looking deeper into different types of breast cancer. If we understand the reasons for the development of cancers better, we can find better ways for fighting them. It would be interesting to continue studying the gene expression differences between basal- and non-basal-like breast cancers. In this project we only looked into a few most differentially expressed genes, but a more thorough analysis is required for determining the underlying differences between these two types of tumors.

5.2 Summary

In this project we looked into the usefulness of centrality in biological data analysis. We wrote the software for calculating average path lengths for large networks and then used it in three different projects.

First we looked into p53 pathway. p53 network regulates the cell cycle and thus plays an important part in the biology of cancer. We calculated the centrality measures for each protein and sorted them accordingly. We found that the most central genes in the network are indeed crucial in the development of cancers and mutations in any of them can have serious consequences for the natural progression of cell proliferation and apoptosis. p53 is the most central protein in the network and its importance as a tumor suppressor protein is widely acknowledged. Cdk2 and CycA are the second and third most central proteins in the network. Cdk2, or cyclin-dependent kinase 2, is essential for the G1/S transition in cell cycle. This protein associates with and is regulated by the regulatory subunits of the complex including cyclin E or A. Binding with Cyclin A (CycA) is required to progress through the S phase. S phase starts when DNA synthesis commences and when it is complete, all of the chromosomes have been replicated. It is easy to see the importance of these three proteins in the context of cell cycle.

Secondly, we analysed a gene expression dataset of different types of breast cancers and normal breast cells with the focus on expression differences between different groups. We used clustering for splitting the data into more manageable partitions and calculated the ratios between the averages of different types of breast cancers to locate the clusters with the most differentially expressed probes. We then ran the centrality algorithm on those clusters and showed that in most cases the most central probes distinguish different types of cancers best. We also analysed the expression levels of the proteins from the p53-network in this context and found that many of these genes are differentially expressed in the tumors in comparison to normal breast cells.

Finally we used clustering and centrality for finding the best compounds for treating breast cancer according to the *in vitro* cancer cell growth inhibition data from NCI. We found that enhydrin and kedarcidin would make for perfect trial drugs as they only target specific cancer cell lines. This implies that there might be considerably fewer side effects when using these compounds for cancer chemotherapy.

As a conclusion, we have shown with this work that centrality can be useful in many different applications. Its significance in protein-protein interaction networks is easy to see, wheres in drug vs. cell line and gene expression based networks we might have to look deeper to understand the meaning of centrality. However, we have displayed here that applying centrality on large datasets can help us significantly with finding the information we are looking for.

References

[1] Albert, R., Jeong, H. & Barabási, A.-L.: *Error and attack tolerance of complex networks*, Nature 306, p.378-382, 2000

[2] De Azua, I.R. et al.:*RGS4 is a negative regulator of insulin release from pancreatic β-cells in vitro and in vivo*, PNAS, vol. 107, 27 April 27 2010

[3] Barabási, A.-L. & Albert, R.: *Emergence of scaling in random networks*, Science 286, p.509-512, 1999

[4] Barabási, A.-L. & Oltvai, Z. N.: *Network Biology: Understanding the cell's functional organization*, Nature Reviews, Vol 5, p.101-113, 2004

[5] Chung, F. & Lu, L.: *The average distances in random graphs with given expected degrees*, Proc. Natl Acad. Sci. USA 99, p.15879-15882, 2002

[6] Cohen, R. & Havlin, S.: *Scale-free networks are ultra small*, Phys. Rev. Lett. 90, 058701, 2003

[7] Dartnell, L. et al: *Robustness of the p53 network and biological hackers*, FEBS Letters. 579. p. 3037-3042, 2005

[8] Ding, L., Hedge, A.N.:*Expression of RGS4 splice variants in dorsolateral prefrontal cortex of schizophrenic and bipolar disorder patients*, Biol Psychiatry, 15 March 2009

[9] Emilsson, L. et al.:*Low mRNA levels of RGS4 splice variants in Alzheimer's disease: association between a rare haplotype and decreased mRNA expression*, Synapse, 1 March 2006

[10] Feltenstein, M.W. et al.:*Anti-inflammatory and anti-hyperalgesic effects of sesquiterpene lactones from Magnolia and Bear's foot*, Pharmacology Biochemistry and Behavior, Volume 79, Issue 2, p.299-302, October 2004

[11] Gail, P.R. et al.:*Breast and prostate cancer: more similar than different*, Nature Reviews, Volume 10, p. 205-212, March 2010

[12] Giaever, G. et al: *Functional profiling of the Saccharomyces cerevisiae genome*, Nature 418, p.387-391, 2002

[13] Giot, L. et al: *A protein interaction map of Drosophila melanogaster*, Science

302, p.1727-1736, 2003

[14] González-Barrios, J., Quiroz, A.:*A clustering procedure based on the comparison between the k nearest neighbors graph and the minimal spanning tree*, Statistics & Probability Letters 62, 1, p.23-34, 2003

[15] Gopalkrishnan, R.V., Lam, E.W.-F., Kedinger, C.:*The p53 Tumor Suppressor Inhibits Transcription of the TATA-less Mouse DP1 Promoter*, J Biol Chem. 273. p. 10972-10978, 1998

[16] Green, K.N., LaFerla, F.M.:*Linking Calcium to Aβ and Alzheimer's Disease*, Neuron, Volume 59, Issue 2, 190-194, 31 July 2008

[17] Inostroza-Ponta, M.:*An Integrated and Scalable Approach Based on Combinatorial Optimization Techniques for the Analysis of Microarray Data*, PhD thesis, University of Newcastle, Australia, 2008

[18] Inoue, A. et al.:*Antifungal melampolides from leaf extracts of Smallanthus sonchifolius*, Phytochemistry, Volume 39, Issue 4, Pages 845-848, July 1995

[19] Inoue, A. et al.:*Hypoglycemic activity of leaf organic extracts from Smallanthus sonchifolius: Constituents of the most active fractions*, Chemico-Biological Interactions, Volume 185, Issue 2, Pages 143-152, 29 April 2010

[20] Jeong, H. et al: *The large-scale organization of metabolic networks*, Nature. 407. p. 651-654, 2000

[21] Jeong, H., Mason, S. P., Barabási, A.-L. & Oltvai, Z. N.: *Lethality and centrality in protein networks*, Nature 411, p.41-42, 2001

[22] Khosravi, R. et al:*Rapid ATM-dependent phosphorylation of MDM2 precedes p53 accumulation in response to DNA damage*, Proc Natl Acad Sci U S A. 96. p. 14973-14977, 1999

[23] Levine, A.J., Hu, W., Feng, Z.: *The p53 Pathway: what questions remain to be explored*, Cell Death and Differentiation. 13. p. 1027-1036, 2006

[24] Li, S. et al: *A map of the interactome network of the metazoan, C. elegans*, Science 2, Jan 2004

[25] Liang, W.S. et al.:*Altered neuronal gene expression in brain regions differentially affected by Alzheimers Disease: A reference data set*, Physiol. Genomics 33: 240-256, 2008

[26] Lill, N.L. et al:*Binding and modulation of p53 by p300/CBP coactivators*, Nature. 387. p823-827, 1997

[27] Liu, X. et al.:*Enzyme-inhibitor-like tuning of Ca channel connectivity with*

calmodulin, Nature 463, 968-972, 18 February 2010

[28] Maslov, S. & Sneppen, K.: *Specificity and stability in topology of protein networks*, Science 296, p.910-913, 2002

[29] Miller, R.T.:*Immunohistochemistry in the Recognition of "Basal-like" or "Basaloid" Breast Carcinoma*, ProPath, 2005

[30] Pastor-Satorras, R., Smith, E. & Sole, R.: *Evolving protein interaction networks through gene duplication*, J. Theor. Biol. 222, p.199-210, 2003

[31] Pusztai, L. et al: *Molecular Classification of Breast Cancer: Limitations and Potential*, The Oncologist. 11. p. 868-877, 2007

[32] Rae, J.M. et al.:*MDA-MB-435 cells are derived from M14 Melanoma cells - a loss for breast cancer, but a boon for melanoma research*, Breast Cancer Res Treat 104:13-19, 2007

[33] Ravetti, M.G. et al.:*Uncovering Molecular Biomarkers That Correlate Cognitive Decline with the Changes in Hippocampus' Gene Expression Profiles in Alzheimers Disease*, PLoS ONE, 13 April 2010

[34] Richardson, A.L. et al.:*X chromosomal abnormalities in basal-like human breast cancer*, Cancer Cell, Volume 9, Issue 2, p.121-132, February 2006

[35] Risbridger, G.P. et al:*Breast and prostate cancer: more similar than different*, Nature Reviews. 10. p.205-212, 2010

[36] Rother, K. et al:*p53 downregulates expression of the G1/S cell cycle phosphatase Cdc25A*, Oncogene. 26. p. 1949-1953, 2007

[37] Saito, T. et al.:*Somatostatin regulates brain amyloid β peptide Aβ42 through modulation of proteolytic deregulation*, Nat Med 11: 434439, 2005

[38] Shoemaker, R.H.:*The NCI60 human tumour cell line anticancer drug screen*, Nature Reviews Cancer 6. p. 813-823, October 2006

[39] Soussi, T.:*The p53 pathway and human cancer*, British Journal of Surgery. 92. p. 1331-1332, 2005

[40] Valenzuela, M.T. et al:*PARP-1 modifies the effectiveness of p53-mediated DNA damage response*, Oncogene. 21. p. 1108-1116, 2002

[41] Zein, N. et al.:*Selective proteolytic activity of the antitumor agent kedarcidin*, Biochemistry, Vol. 90, pp. 8009-8012, September 1993

[42] `http://www.cancercouncil.com.au/html/policyaction/reports/downloads/costofcancer_summary.pdf`

Retrieved May 20, 2011

[43] http://dtp.nci.nih.gov/docs/misc/common_files/mda-mb-435.html

Retrieved May 20, 2011

[44] http://dtp.nci.nih.gov/docs/misc/common_files/NCI-ADRres.html

Retrieved May 20, 2011

[45] http://en.wikipedia.org/wiki/Anthracycline

Retrieved May 20, 2011

[47] http://en.wikipedia.org/wiki/Dijkstra's_algorithm

Retrieved May 20, 2011

[48] http://en.wikipedia.org/wiki/Breast_cancer

Retrieved May 20, 2011

[49] http://en.wikipedia.org/wiki/Myc

Retrieved May 20, 2011

[50] http://en.wikipedia.org/wiki/Ras_subfamily

Retrieved May 20, 2011

[51] http://en.wikipedia.org/wiki/Abl_gene

Retrieved May 20, 2011

[52] http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene&cmd=Retrieve&dopt=Graphics&list_uids=993

Retrieved May 20, 2011

[53] http://centibin.ipk-gatersleben.de/examples.php

Retrieved May 20, 2011

[54] http://www.ncbi.nlm.nih.gov/projects/geo/gds/gds_browse.cgi?gds=2250

Retrieved May 20, 2011

[55] http://dtp.nci.nih.gov/docs/cancer/cancer_data.html

Retrieved May 20, 2011

[56] http://en.wikipedia.org/wiki/Minimum_spanning_tree

Retrieved May 20, 2011

[57] http://en.wikipedia.org/wiki/K-nearest_neighbor_algorithm
Retrieved May 20, 2011

[58] http://en.wikipedia.org/wiki/Pearson_correlation
Retrieved May 20, 2011

[59] http://en.wikipedia.org/wiki/Spearman_correlation
Retrieved May 20, 2011

[60] http://en.wikipedia.org/wiki/Euclidean_distance
Retrieved May 20, 2011

[61] http://dtp.nci.nih.gov/docs/misc/common_files/NCI-ADRres.html