Prognostic signatures for Cancer: A computational biology approach

Vasantha Kumar R*. Aditya Gupta**
**BS19B032 - IIT Madras - e-mail: bs19b032@smail.iitm.ac.in.*
***BS18B001 - IIT Madras - e-mail: bs18b001@smail.iitm.ac.in.*

Abstract

Cancer has become one of the deadliest disease in world. According to the World Health Organisation(WHO) reports of 2020, cancer is the reason for most deaths, reaching nearly 10 million. The most common cancer, especially among women is breast cancer and then ovarian cancer. Therefore, in this paper we try to find ways using computational biology for detection of cancer related genes. Prognosis has become an important part in medical biology, which helps in finding symptoms, duration of infection and early detection of diseases. Here, we plan to find the prognostic signatures from gene expression data of breast cancer, which can then be used as biomarkers, to detect disease. This is done by creating a network module using the gene networks data obtained. We hope to find these markers, which might later help in treatment and detection of cancer. The paper mainly describes a simple and rapid way to combine expression data of genes that are disease specific with a static protein functional interaction network in order to identify candidate prognostic network modules.

Keywords: *Prognostic, protein network, breast cancer, biomarkers*

1. Introduction

One of the main goals of genomic application is to find ways to distinguish patient subtypes which are otherwise is indistinguishable. This is done by identifying several clinically relevant and important biomarkers. This would then help the clinicians to take judicious decisions for providing aggressive therapy for patients who tend to show a higher chance of developing/acquiring aggressive diseases. Such tests are also quite helpful in choice of therapies which are most likely to benefit these patient groups [1,2,3].

The advancements in network and systems biology such as studies dedicated to global relationships between human diseases, genes and interactome models, show how human diseases/ disorders can be thought of as perturbations in cellular networks which are strongly interconnected. Human diseases/ disorders being highly dependent on cellular networks and their perturbations has resulted in emergence of global disease maps, such as gene-disease associations collected in the OMIM database (Goh et al., 2007).

Interactome is a network of gene and protein interactions occurring in cells. When various interactome networks are linked/integrated with networks depicting the functional relationships, it helps to reveal potential genes involved in cancer. Together including with genetic interactions and physical ones would yield a breast cancer network model, which can be then used to predict cancer susceptibility as well as modifier genes (Pujana et al., 2007) [4].

According to researchers, multi gene expression signatures present a far clearer picture than single gene expression, in understanding and decoding the abnormalities between phenotypes with respect to perturbations in biological networks. In recent times, analytical approaches based on gene expression networks and expression data sets help to get insights and predictions of clinical outcomes of various types of cancers. So, here we are interested in finding gene module with many genes that aid in cancer research.

According to WHO, breast Cancer (2.26 million cases) [6] is considered to be one of the leaders in cancer death numbers all round the world, perhaps the most common cancer among women. Owing to the heterogeneity in clinical presentation and progression, breast cancer is considered one of the favourite models for the development and testing of multi-gene prognostic signatures.

1.1 Objectives

In this project, our main objective is to find gene markers in a human protein interaction network that would affect the survival of patients. This is done using the microarray mRNA data of cancer patients.

Then, we would like to validate the found gene module using Kaplan-Meier survival analysis, to understand how the found gene module affects the survival of pateints.

2. Materials and Methods

2.1 Protein Interaction Network

To construct the protein interaction network, we used the network module created by *Guanming Wu [6].* We used this network because, it contains interactions of diverse pathways like apoptosis and cell-cycle progression, which are important in cancer research. Also, this network also contains predicted interactions, which might help in finding new gene markers.

2.2 Adding weight to network

Next, to make this network more related to cancer, we introduced weights to the edges. These weights are Pearson Correlation Coefficients(PCC) between interacting genes. To calculate PCC, we used the algorithm developed by *Jamil Najafov, Ayaz Najafov called GECO [7]. From GECO website, we downloaded the R source code for the algorithm and modified it to find PCC values based on CCLE (Cancer Cell Line Encyclopaedia) database.*

2.3 Clustering of network

For efficient working on the network, we formed clusters in the network using the MCL or Markov clustering algorithm [8]. We downloaded the algorithm from the author's website. As it is a large network, to reduce the size of clusters, inflation was set to 20. To find clusters for efficient usage, we set two thresholds on the clusters: total genes(n) > 7 and average edge weight > 0.25.

Figure 1: Processing steps for the protein interaction network

2.4 Breast Cancer Dataset

The Breast Cancer dataset we used in the module was downloaded from GEO as soft format files [9]. The dataset is GSE1456, which contained gene expression data of 159 cancer patients. Then, the probes with affymetrix ids in dataset were matched to genes using online bioinformatics tool GEO2R and then normalized it [10].

2.5 Finding network module related to cancer data

Next, using the cancer data, we calculated the total gene expression for each module. Then, we also calculated differently expressed genes in the data based on survival of patients. Now, we went for a linear search for a network module matching with these conditions.

2.6 Kaplan-Meier survival analysis

Then, to validate the found network module, we did a Kaplan-Meier survival analysis on the cancer dataset, based on the expression of the found module on each patients. We downloaded the R package for survival analysis and modified it [11].

3. Results

The final protein interaction network when plotted contains a total of 9450 genes with 181691 interactions. The average degree of genes in the network was about 38. This network is initially unweighted, moreover, the network is general to all cells in the body. So, to make this network specific to cancer, we added PCC value between genes as edge weights. The absolute values of PCC is taken, as weight cannot be negative. Also, for some variant genes in the network we were unable to find PCC values, so we considered weights as 0. The final network is then visualised in the cytoscape(fig.2).

Now, this weighted network is subjected to MCL clustering to produce network clusters which are closer in topology, also, has high gene expression among them. The inflation was set to 20 to produce considerably small clusters. The initial total clusters formed were 6711. Then to find efficient clusters, at the threshold of n>7, we got a total 95 clusters. Then these were subjected to second threshold of average weights > 0.25, which yielded final filtered clusters of 42.

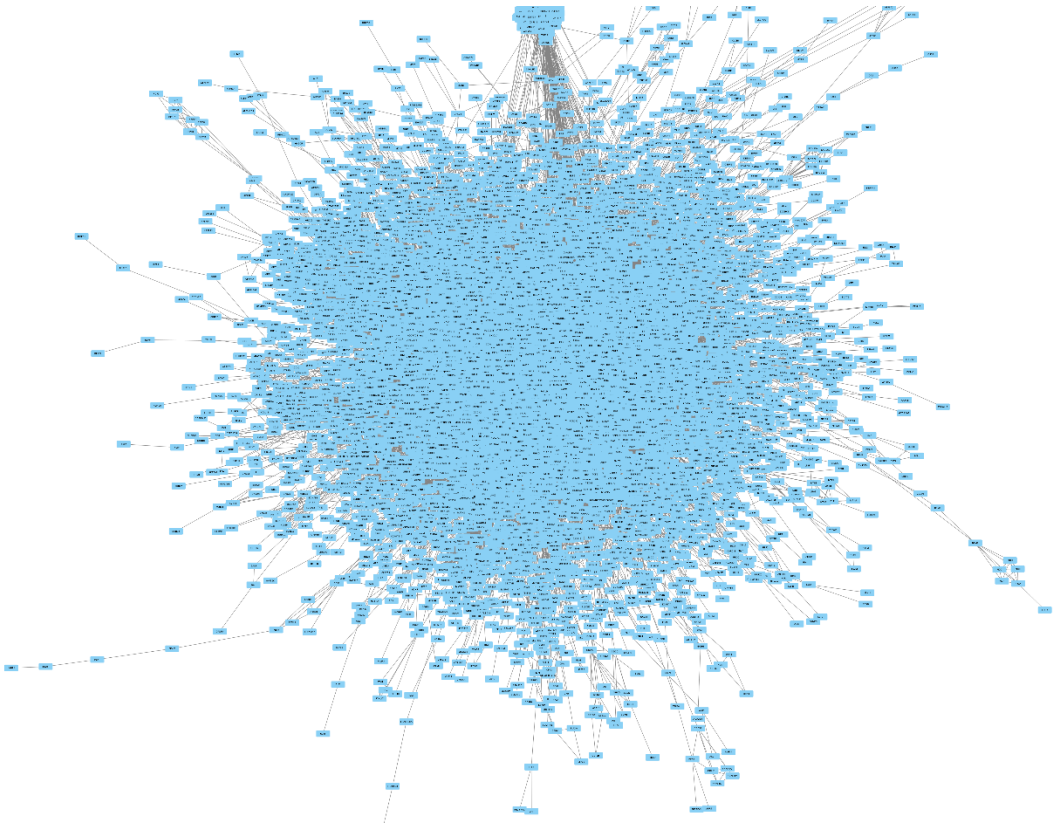

figure 2: The protein interaction network in cytoscape

The dataset obtained contains the survival time of patients in years, relapses and survival relapses. It also contained microarray expression data for each gene in all patients. Then, by processing the obtained data, we matched probes to gene. Also, we found the differentially expressed genes in the dataset according to the survival statistics of patients. The top 5 such genes are listed in table1.

The gene expression value of each cluster for each patient is then stored in a 2D matrix. Then we did a linear search to find the network modules which are highly expressed across all patients. Also, network modules with differentially expressed genes were also ranked separately. Now, combining these two results we obtained the final network module which significantly affects the survival of the breast cancer patients. The final module contained a total of 21 genes [*PIK3R1, STAT5B, TINF2, SNAI1, GNAT2, SPDEF, OR5A2, MLL3, ATP4A, SEC11A, RAD21, C1QTNF5, CELSR1, MARS, DTX4, PRPF3, FUT1, SP110, LRAT, REXO2, SERPINB3*], we named it module X.

Now, to validate the module X to find whether it affects the survival of cancer patients, we did Kaplan-Meier survival. It is a univariate method which is used to predict the survival probability, based on the specified variables. First, we calculated the gene expression of module X across all patients and ranked them accordingly. Next, we found the median for this data. Now, the patients were split into two groups: patients above median value and below. We then plotted the curve for two sets in same graph for comparison purposes (fig 3).

From figure 3, it is clear that, high expression of module X genes, results in low survival of breast cancer patients. Also, the plotted graph has very low p value of less than 0.001.

ID	Adj.P value	P value	t	Gene
212549_at	0.00493	0.000000221	- 5.4139547	STAT5B
217957_at	0.02493	0.00000287	4.8518665	CFAP20
209285_s_at	0.02493	0.00000438	- 4.7551947	FAM208A
212070_at	0.02493	0.00000448	4.7504831	ADGRG1
215011_at	0.02898	0.0000065	4.663706	SNHG3

Table 1: Top 5 differentially expressed genes in the dataset based on survival of patients

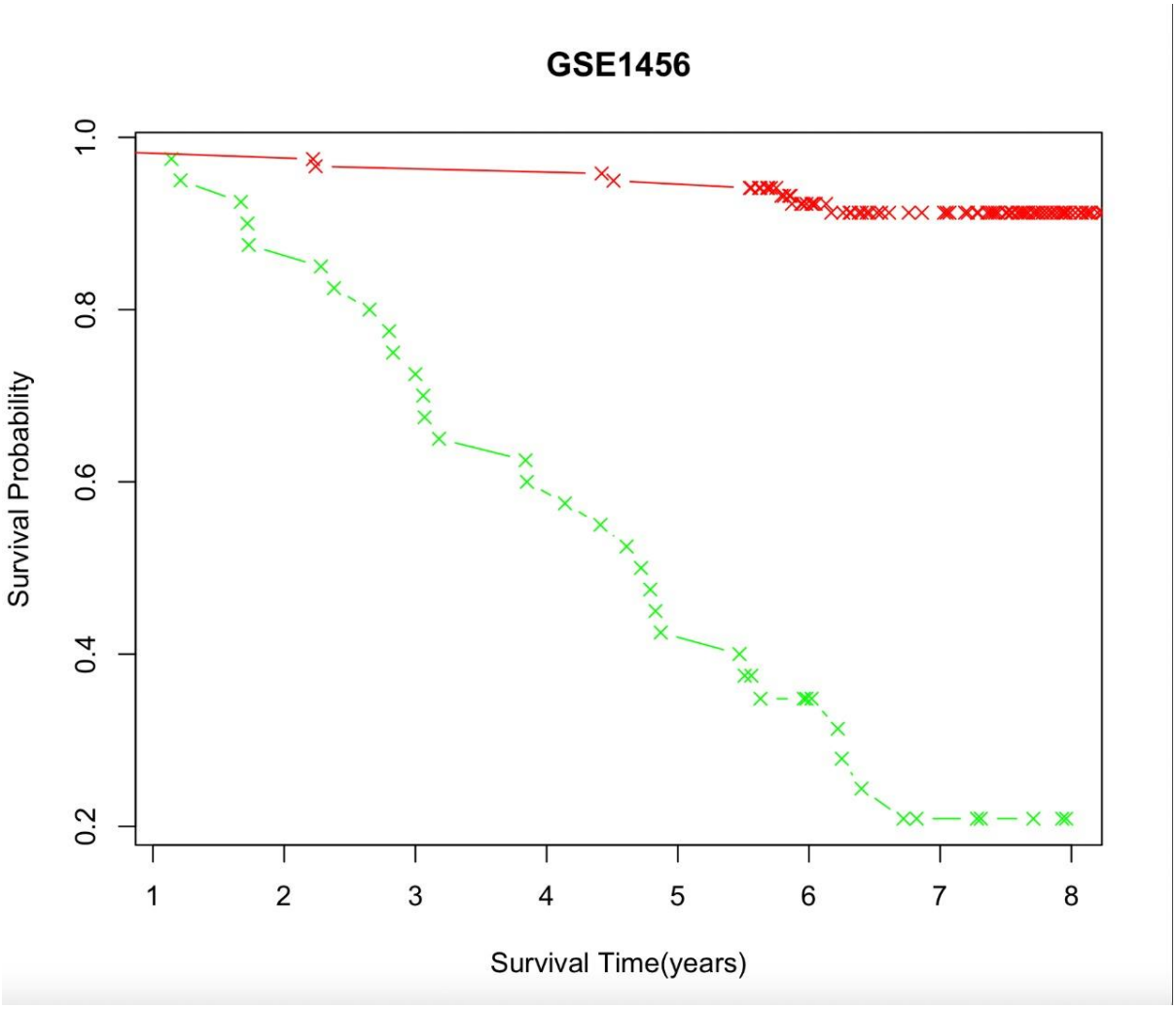

Figure 3: Kaplan-Meier survival plot based on module X

Now, to understand why the module X genes affect the survival of the patients, we studied the functions of the genes in cell. There the highly differentially expressed gene STAT5B (table 1), plays a significant role as a transcription factor, which gets activated by growth factors. It acts as a signal transducer and involves in various processes like apoptosis and adult mammary gland development [12]. So, it is clear that STAT5B is a highly important gene in breast cancer analysis.

Other genes like TINF2, which affects telomeres function, thus directly interpreting cell divisions. The CELSR1 gene mediates contact communications, RAD21 gene is significant in M phase of cell cycle and REXO2 gene provides resistance to cell death. Therefore, as the module X genes plays a significant role in cell cycle, contact inhibitions, these genes are highly related to survival in breast cancer patients.

4. Discussion

In this project, we did a simple and rapid procedure to combine disease specific gene expression data along with a static protein functional interaction network, in order to identify candidate prognostic network modules and genes. The disease we focused was the breast cancer. Though this we managed to find a module of genes, which involves significantly in cancer related activities of cell such as contact mediation, mitosis and signal transduction.

But, a problem with the module X genes is the not all genes in the cluster directly affects the cancer characteristics of cell, but, still they involve in cell cycle related activities. As low p value in Kaplan-Meier survival analysis, indicate that overall, the module X has a robust involvement in breast cancer.

An important problem, we faced in this project was the clustering of network. The MCL algorithm we used for clustering produced inconsistent results for us, leading to formation different clusters, at each time. Next, drawback is about the diverseness of the dataset used. The dataset we used GSE1456, contains patient details only from Stockholm, Sweden. Thus due to the genetic factors, the significance of this result across others countries might be less.

5. Conclusions

As cancer cases in world is at a steady rise, cancer related researches are important. Here, using computational biology, we produced a module of genes which has significant effects in breast cancer patients. Perturbations of biological networks within cells is crucial to help interpret how genome variations relate to phenotypic differences. This might help us in identifying cancer phenotypes easily and early, which would then aid in treatment and cure for cancer.

A module X containing genes [PIK3R1, STAT5B, TINF2, SNAI1, GNAT2, SPDEF, OR5A2, MLL3, ATP4A, SEC11A, RAD21, C1QTNF5, CELSR1, MARS, DTX4, PRPF3, FUT1, SP110, LRAT, REXO2, SERPINB3] was identified and validated whose high expression resulted in low survival of breast cancer patients. As low p value in Kaplan-Meier survival analysis, indicate that overall, the module X has a robust involvement in breast cancer.

References

1. Kaufmann M, Pusztai L, Members BEP: Use of standard markers and incorporation of molecular markers into breast cancer therapy: Consensus recommendations from an International Expert Panel. Cancer 2011, 117:1575-1582.

2. Subramanian J, Simon R: Gene expression-based prognostic signatures in lung cancer: ready for clinical use? J Natl Cancer Inst 2010, 102:464-474.

3. Brugger W, Triller N, Blasinska-Morawiec M, Curescu S, Sakalauskas R, Manikhas GM, Mazieres J, Whittom R, Ward C, Mayne K, Trunzer K, Cappuzzo F: Prospective molecular marker analyses of EGFR and KRAS from a randomized, placebo-controlled study of erlotinib maintenance therapy in advanced non-small-cell lung cancer. J Clin Oncol 2011, 29:4113-4120

4. Vidal M, Cusick ME, Barabási A-L: Interactome networks and human disease. Cell 2011, 144:986-998.

5. WHO: Cancer.. [http://www.who.int/mediacentre/factsheets/fs297/en/].

6. A Human Functional Protein Interaction Network and Its Application to Cancer Data Analysis - Guanming Wu, Xin Feng and Lincoln Stein.

7. GECO: gene expression correlation analysis after genetic algorithm-driven deconvolution by Jamil Najafov, Ayaz Najafov.

8. van Dongen S: Graph Clustering by Flow Simutation. PhD thesis University of Utrecht; 2000.

9. Gene Expression Omnibus (GEO).. [http://www.ncbi.nlm.nih.gov/geo/].

10. GEO2R (https://www.ncbi.nlm.nih.gov/geo/geo2r)

11. STDHA - Statistical tools for high-throughput data analysis

12. The STAT5b pathway defect and autoimmunity Takahiro Kanai†, Jennifer Jenks† and Kari Christine Nadeau*