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.*

First Project Report

Cancer has become one of the deadliest disease in our world. Also, the rate of cancer is ever increasing. According to the, Global Cancer Statistics, 2020, there are 19.3 million new cancer cases, out of which almost half, 10 million death cases are reported [1]. Among, females breast cancer and ovarian cancer have become most common cancer detected. So, in this project our idea is to find biomarkers in genes, which would help in the survival of cancer patients.

In our step towards the project, we wanted to create a network of gene interactions, which would consist of protein pair relationships among different proteins. For this we used the functional human protein interaction network proposed by Wu G, Feng X and Stein L, for cancer application [2]. The downloaded file was consisted of gene names and its UniProt accession numbers. The main reason we selected this network is that, apart from the curated pathways of genes, this also contains, predicted gene interactions using Bayer's classifier. Moreover, this network has common patterns with mechanisms in cancer biology.

Next, using the obtained network, we plotted the gene interactions. The final network consists a total of 9450 genes, 181691 interactions, with the average degree of 38.4531. But this network is unweighted, that is, the strength of interactions is not known.

So, we found the Pearson Correlation Coefficient(PCC) for each pair of genes, and added their absolute value, as the edge weight. But, finding PCC for the genes was challenging as we cannot use a mathematical formula. So, for finding PCC, we used GECO – it is a gene expression correlation analysis using deconvolution [3]. It is actually an app.; from which we could find all the correlations of given two genes. We downloaded the R source codes for the app given in their website. Then we modified it to produce the PCC for all the 181691 interactions in our network. The database used to correlate was CCLE (Cancer Cell Line Encyclopaedia). There were some variant proteins in the network for which we cannot correlate, so we assumed the PCC of their interactions to be zero.

But, the problem is that our network was too dense with many interactions. So, we decided to form clusters of our network using the highly efficient MCL clustering algorithm [4]. We downloaded the MCL package from the original website, and configured it to form clusters. We filtered the clusters formed with threshold average edge weight = 0.25 and number of genes = 8.

Now, the challenging step in our project is to obtain the breast cancer and ovarian cancer datasets. The datasets used in our project article [5], was from Netherlands Cancer Institute(NKI) for which we do not have access. So instead we downloaded the breast cancer and ovarian cancer datasets from GEO(Gene Expression Omnibus) which is managed by the NCBI. All the datasets downloaded was in SOFT format [6]. Our primary dataset for breast cancer consist survival relapses of a total 159 patients, along with their mRNA gene expressions which were affected.

Now, our next step is to map the probes to genes in the given dataset. Then, we would do the module search to find the biomarkers. Later, we would also like to plot a survival analysis and the p-values.

Figure 1: Plot of the network

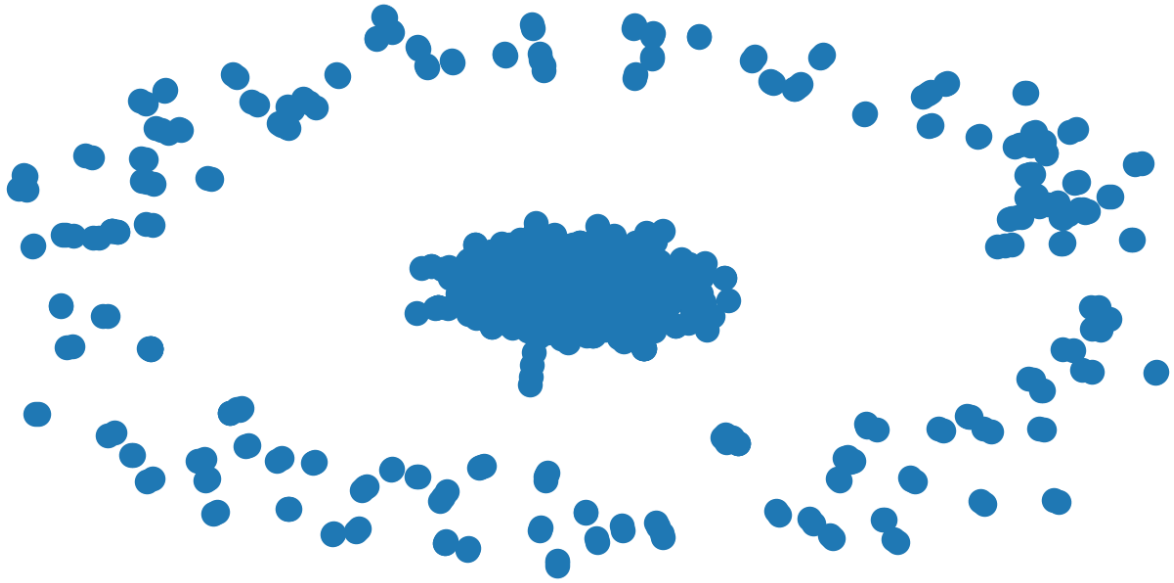

Figure 2: GECO – App to find PCC

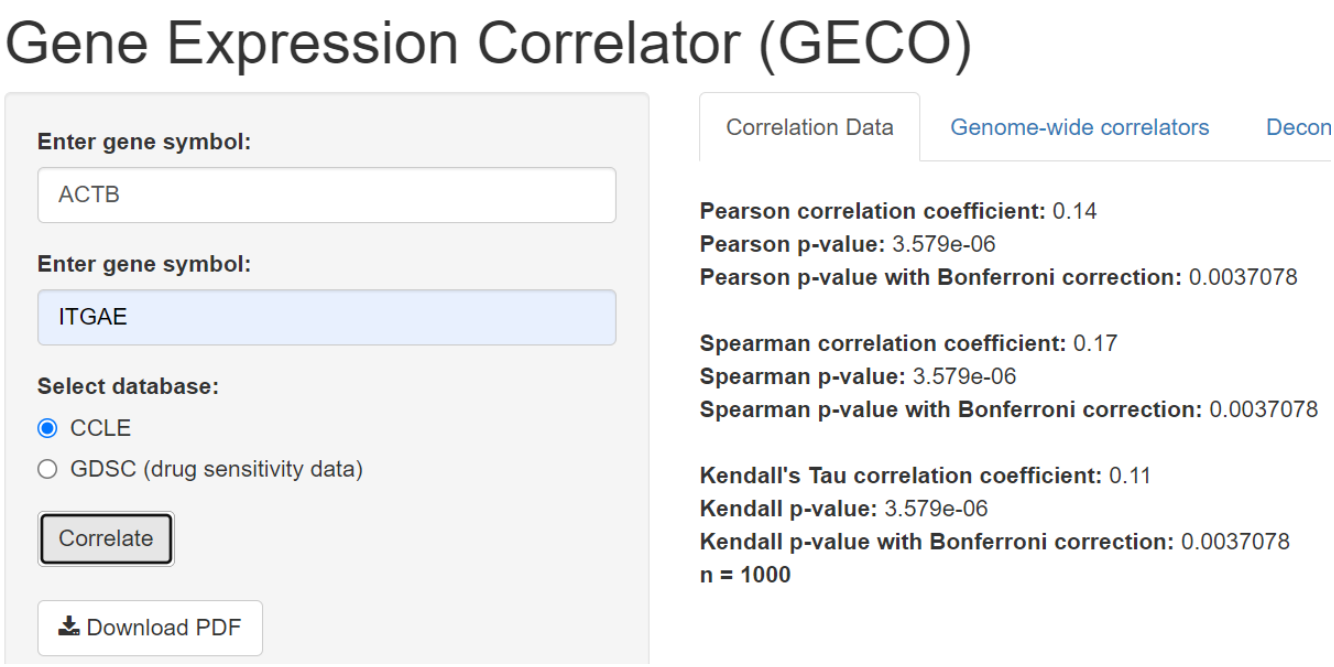

Figure 3: Breast Cancer dataset

ID_REF	RELAPSE	SURV_RELAPSE	DEATH	DEATH_BC	SURV_DEATH	SUBTYPE ELSTON
X027JO	1	3.82	1	4.14	No Subtype	2
X350JO	0	8.15	0	8.15	Luminal B	3
X028JA	0	2.22	1	2.22	Luminal A	1
X126AS	0	8.23	0	8.23	No Subtype	2
X005JO	0	5.55	1	5.55	Luminal A	NA
X045OL	0	8.3	0	8.3	Luminal B	3
X041LA	0	4.42	1	4.42	Luminal A	2
X229LA	0	8.07	0	8.07	No Subtype	1
X347JA	0	6.38	0	6.38	Basal	3
X122FO	0	8.13	0	8.13	ERBB2	3
X054BO	1	4.08	1	5.51	Basal	NA
X204FA	0	8.11	0	8.11	No Subtype	2
X356AL	1	3.47	1	5.47	Luminal B	3
X089HO	1	0.56	0	6.82	Luminal B	3
X215HE	1	5.53	1	6.25	Normal Like	2
X223LI	0	7.71	0	7.71	No Subtype	2
X055SV	1	4.44	1	6.22	Luminal A	2
X380MO	1	1.39	1	3.06	Basal	3
X111MA	0	5.87	1	5.87	Basal	3
X264LA	0	7.86	0	7.86	Luminal A	2
X101GL	1	2.92	1	3.85	Luminal A	2
X296NY	1	1.09	1	2.65	Luminal B	3
X185NY	1	4.36	1	5.63	Luminal B	3
X057FR	0	7.61	0	7.61	No Subtype	2
X298LA	0	7.55	0	7.55	Normal Like	1
X376FR	1	1.19	1	3.18	Luminal A	2
X164MY	0	8.03	0	8.03	Normal Like	1
X071AV	1	5.16	0	7.31	Luminal A	2
X377DE	0	7.65	0	7.65	Normal Like	3
X166JO	0	7.38	0	7.38	Basal	2
X187OL	0	7.29	0	7.29	Basal	3
X121LE	0	8.03	0	8.03	Luminal A	2
X318KI	0	8.03	0	8.03	Normal Like	1

References:

[1] CA: A Cancer Journal for Clinicians / Volume 71, Issue 3 / page no. – 209 to 249.

[2] Wu G, Feng X, Stein L: A human functional protein interaction network and its application to cancer data analysis. Genome Biol 2010, 11:R53.

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

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

[5] A network module-based method for identifying cancer prognostic signatures - Wu and Stein Genome Biology 2012, 13:R112.

[6] https://www.ncbi.nlm.nih.gov/geo/query acc=GSE1456