Applying Topological Methods to Gene Expression Analysis

Conrad De Peuter, BS, Vijayaraghavan Balaji, B.Tech

Columbia University, New York, NY, USA

Abstract

Finding differentially expressed genes from cancer microarrays is a common technique used in determining the causes of cancer, but extracting meaning from this set of genes is sometimes not straighforward. Regular clustering methods routinely group genes that are not biologically related, thus it is hard to extract information from such analyses. We propose using a topological data analysis method known as Mapper in conjunction with data on gene functions as well as the pathways where they operate to gain biological insight from Gene Expression Analysis. We propose a method which will identify meaningful subgroups of cancer which will help identify prognosis and treatment.

1. Introduction

The Mapper method, introduced by Carlsson et. al. in 2007 [1] formulated a novel topological approach to analysis of high-dimensional data sets. The idea is based on partial clustering of the data combined with a set of biologically meaningful functions on the data to reduce its dimensionality. With a well defined set of filters, Carlsson et. al. were able to show that relationships among data points persisted throughout the dimensional reduction that did not persist in regular clustering algorithms. Applying the Mapper method to breast cancer microarray data in 2011, and introducing their own variation of this method, termed Progression Analysis of Disease (PAD) Nicolau et al. [2] were able to identify certain breast cancers with unique gene profiles, and discovered that two of these cancer types had 100% survival rates and no metastasis. We hope to use Mapper to identify biologically significant clusters of genes in the same manner.

In our research we would like to apply the Mapper/PAD method to other cancer microarray data and see if we can identify subgroups of cancers with common characteristics which would help determine prognosis and treatment methods. We would like to see if the Mapper method gives a different clustering versus standard clustering methods, and if so we would like to investigate differences in these clusters. Building on Nicolau et. al. we also intend on looking at the pathways where these differentially expressed genes operate to find significant relationships between the significant pathways.

2. Methods

2.1. Data Sources

We used two datasets from the Gene Expression Omnibus. GDS5437 [3] compares 14 lung tissue samples of gene expression of healthy mice versus mice with non-metastatic and metastatic breast cancer tumors. These samples consisted of 5 healthy mice, 5 with non-metastatic tumors, and 4 with metastatic tumors. The original study for the dataset was looking at the over-expression of the G-CSF glycoprotein in the lung tissue of mice with metastatic breast cancer, and showed that in the lung tissue of these mice only those with metastatic breast cancer showed over-expression of G-CSF. Because the samples in this data are from lung tissue of mice with breast cancer, the cancer may have not spread to the lungs and we may not see any significant expression in relevant cancer genes, because of this we are also looking at a second dataset. GDS1439 [4] compares 19 samples of prostate cancer tumors in humans which are benign (6), clinically localized (7), and metastatic (6). The original study for this dataset looked to identify which proteins were altered in the different states of cancer and tried to identify a proteomic progression signature in prostate tumors.

2.2. Preliminary Methods

Initially we fit a linear model to each of the datasets to identify which genes were differentially expressed. After finding these genes we performed two types of cluster analysis on the data. We compared the clusters produced by hierarchical clustering with those produced by the Mapper/PAD method and compared the clusters produced by these algorithms. The clusters produced by Mapper are different than the hierarchical clusters, and we hypothesize that there are meaningful characteristics in these clusters which will help identify prognosis/treatment.

2.3. Proposed Methods

After gathering clusters of genes using Mapper, we plan on using the Gene Ontology database to map genes to their biological function, and see if the genes clustered together by Mapper have related functions. We plan on using a chi-squared test to determine whether any biological function is overrepresented in a group. In addition, we will use the KEGG Mapper to determine the pathways where genes operate and do similar analysis to see if we have found any significant pathways.

2.4. Evaluation Strategy

After completing the analysis detailed above we plan on evaluating the results in two ways. We have compiled a gold standard of known significant genes associated with prostate (38 genes [6]) and breast cancer (26 genes [5]) collected from the Genetics Home Reference, and will group them together by biological function. We will then verify our hypothesis by seeing whether our method groups the genes in a similar way to the gold standard, and whether this grouping could have happened by chance. Secondly, we will look at more microarray data for another form of cancer and see if our proposed method can group genes of biological significance in that data as well. If so this would be a good indication that our method is not only viable, but also can be applied across a wide range of cancer data.

3. Preliminary Results

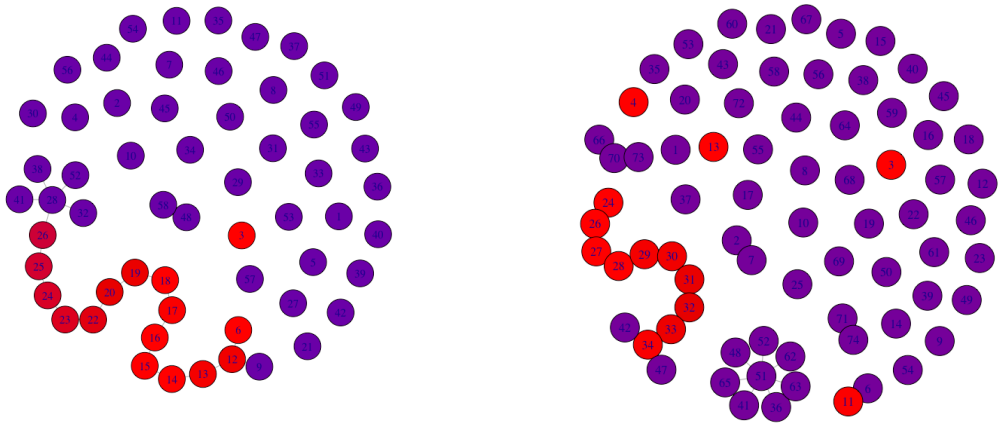

Figure 1: Mapper output for GDS5437 (left) and GDS1439 (right). Shade of vertex corresponds to percentage of DEG in vertex

Using Mapper for GDS5437 we found that 93% of the DEG ended up in a set of vertices linked together. Using hierarchical clustering and cutting the tree so that there were the same number of clusters as the Mapper algorithm, the top 9 clusters only had 75% of the DEG. In GDS1539, Mapper grouped the top 94% of DEG in the 10 joint clusters, while only 86% of them were clustered hierarchically. In addition, using the Biological Homogeneity Index introduced by Datta et al. [8] we were able to confirm that the clustering produced by Mapper on GDS5347 was more biologically homogenous than both k-means and hierarchical clustering.

Table 1: BHI results for GDS 5437

Mapper	Hierarchical	K-Means
.4	.37	.28

4. Discussion

4.1. Summary

We are looking to analyze if clustering gene expression data of cancer samples using Mapper with a biologically significant filter produces more significant clusters than regular methods. The preliminary analysis shows that the clusters obtained have a stronger concentration of differentially expressed genes as compared to hierarchical clustering. The main objective is to identify the significance of these genes being in the same clusters, and what could be the biological implications of the underlying relationships between the genes clustered together. A potential issue with this research is that we may find clustered genes with no related biological function, or already known biological function.

4.2. Anticipated Results

We expect to see clusters of genes in Mapper that don't exist in regular clustering which have related biological functions. We will determine this by looking at whether the genes operate in similar pathways or if they are related to the same biological function. We will evaluate the similarity of these pathways using the methods of Chen et al. [7], who evaluate whether clusters of genes have a significant similar biological function using a Chi. Squared test. If our analysis confirms this then we will see whether that function is known to be associated with the given cancer. If it is not previously known to be associated with the cancer, then we have found a novel area for investigation of this cancer. If we are not able to find any common biological function among the grouped genes then we will look into other datasets, as our approach is easily transferrable. We may also notice that Mapper's clusters are no more biologically significant than the other clustering methods as those may also group functionally related genes. In this scenario we plan on presenting our code base as a tool to assist others in comparing different clustering methods.

4.3. Conclusion

In this research we are looking to apply the Mapper/PAD method introduced by Nicolau et. al. to other cancers, and identify significant cancer subgroups to aide in treatment/prognosis. We are looking to introduce a method to generalize their analysis to more cancer data, and to extend the analysis by identifying biologically significant pathways as well as genes.

5. Citations

[1] Singh G, Memoli F, Carlsson G (2007) in Eurographics Symposium on Point-Based Graphics, Topological methods for the analysis of high dimensional data sets and 3D object recognition, eds Botsch M, Pajarola R (Eurographics Association, Geneva), pp 91?100.

[2] M. Nicolau, A.J. Levine, G. Carlsson Proc. Natl. Acad. Sci. U. S. A., 108 (2011), pp. 7265?7270

[3] Kowanetz M, Wu X, Lee J, Tan M et al. Granulocyte-colony stimulating factor promotes lung metastasis through mobilization of Ly6G+Ly6C+ granulocytes. Proc Natl Acad Sci U S A 2010 Dec 14;107(50):21248-55. PMID: 21081700

[4] Varambally S, Yu J, Laxman B, Rhodes DR et al. Integrative genomic and proteomic analysis of prostate cancer reveals signatures of metastatic progression. Cancer Cell 2005 Nov;8(5):393-406. PMID: 16286247

[5] Breast cancer - Genetics Home Reference. (n.d.). Retrieved November 01, 2016, from https://ghr.nlm.nih.gov/condition/breast-cancer

[6] Prostate cancer - Genetics Home Reference. (n.d.). Retrieved November 01, 2016, from https://ghr.nlm.nih.gov/condition/prostate-cancer

[7] Chen, Gengxin. "EVALUATION AND COMPARISON OF CLUSTERING ALGORITHMS IN ANALYZING ES CELL GENE EXPRESSION DATA." Statistica Sinica 12.1, A Special Issue on Bioinformatics (2002): 241-62. JSTOR. Web. 01 Nov. 2016.

[8] Methods for evaluating clustering algorithms for gene expression data using a reference set of functional classes. Susmita Datta, Somnath Datta BMC Bioinformatics. 2006; 7: 397. Published online 2006 Aug 31. doi: 10.1186/1471-2105-7-397 PMCID: PMC1590054