

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 often not straightforward. Regular clustering methods routinely group genes that are not biologically related, thus it is hard to extract information from such analyses. Using a topological data analysis method known as Mapper in conjunction with data on gene functions as well as the pathways where these genes operate, we find evidence that clusterings produced by Mapper can have more biological significance than regular clustering methods K-Means and Hierarchical Clustering. The functional clustering of these genes may provide insight for novel areas of investigation in determining the causes/prognosis for cancers.

1. Introduction

1.1. Mapper/PAD

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 biologically meaningful filtering function on the data to reduce its dimensionality. The method performs partial clustering on overlapping intervals of the data set. The output is a graph where vertices are the bins within each clustering. Because the intervals are overlapping a data point may end up in multiple clusters. If this happens an edge is drawn between these two clusters in the output graph. 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 a type of breast with unique gene profiles, and discovered that patients with this cancer had 100% survival rates and no metastasis. In this study we intend to provide further evidence of Mapper's ability to produce functionally significant clusterings. According to Carlsson et al.'s original paper on mapper, the goal of the method is to build low-dimensional image of the data set which may indicate areas of interest. Applying this method to cancer data in a similar fashion to Nicolau et. al. we would like to indicate areas of interest for studying two specific forms of cancer, as well as provide a general framework for similar studies in the future.

1.2. Benefits of Discovering Functionally Significant Clusters

The possible benefits of such a technique are clear. [WHAT WERE THE BENEFITS OF NICOLAU, able to find cancer types with 100% survival rates]. Unlike the PAD study, we do not have access to the survival rates of the patients in the data set, we only have the snapshot in time of when the gene expression data was taken. In lieu of this our final goal is not to find specific high or low survival rate cancers, but instead to develop a method which produces areas of interest for further study of these cancers by examining the functions of genes which end up clustered together, as well as pathways where they operate.

2. Methods

2.1. Data Sources

We used two data sets 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 data set 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. We are also looking at a second data set. 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 data set 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. Data Dimension Reduction

Before clustering we reduced the number of genes in each data set, in order to make hundreds of clusterings computationally feasible. We did this by removing genes which did not have a sufficient expression level across enough samples using the *genefilter* package in R. We then fit a linear model to each of the data sets to identify which genes were differentially expressed. GDS1439 produced thousands of differentially expressed genes, so we took the ones with the top 1000 p-values. GDS5437 only produced 46 differentially expressed genes, so we took all of those. For both data sets we reduced each to around 7000 genes.

The crux of the Mapper algorithm is a dimensionality reducing filter function which "reflects geometric properties of the data set" [1]. The topological ideas of the method are based on the idea that if this filtering function is well defined (TODO what does this mean) and significant in the context the data, then the proximity of similar data points should persist throughout this dimensionality reduction. We used the same filter function from the PAD method [2], which isolates the diseased component of every data point/vector by removing the normal component from all points. If two genes perform a similar function in the context of a disease, then they should remain close after the dimensionality reduction. TODO talk about persistence.

After finding these genes and reducing the dimension we performed three types of cluster analysis on the data, using the Biological Homogeneity Index to compare the functional significance of the clusters produced by Mapper, K-Means, and Hierarchical Clustering. To find the optimal clustering of the genes we performed a grid analysis across the parameters for Mapper: number of intervals, overlap between intervals, and number of bins to use when clustering in the intervals. Mapper also takes a point-wise distance matrix as an input, and for this we experimented with both the Pearson and Spearman metrics, seeing slightly better results with Spearman. The output to Mapper is a graph, with vertices as clusters. For each set of parameters we took the number of vertices in the outputted graph, and set that to k for K-Means, and cut the Hierarchical Clustering tree at the level which would produce that many clusters. By looking at the clustering which produced the optimal BHI for Mapper, which happened to be the clustering which had the highest difference in performance compared to the other two methods, we argue that the clustering produced by Mapper is more biologically significant than the other methods.

3. Results

3.1. General Mapper Comparison

Searching through hundreds of parameter combinations we found that for GDS1439, the BHI of Mapper's clustering was noticeably larger than that of either K-Means or Hierarchical Clustering, although not outside of a standard deviation. For the clusterings where hundreds of genes were placed in each cluster there was no noticeable difference, but once the number of clusters was past 100 the BHIs of the Mapper had a noticeable improvement. This result seems somewhat intuitive if you accept the hypothesis that Mapper produces functionally significant clusters; with hundreds of genes in a single cluster there are bound to be many unrelated genes thus a lot of noise. Once the number of genes in a single cluster is of a reasonably small size the opportunity for a method specifically aimed at discovering functional significance to show its worth is realizable. Tables 1 & 2 show how the clusterings' BHI's, for all genes, as well as just the differentially expressed genes in each cluster, changed with respect to number of clusters/cluster size.

While the results were promising for GDS1439, the same cannot be said for GDS5347. In the latter hierarchical clustering produced the best BHI results for the whole clusters, and K-Means did best for differentially expressed genes. While BHI is a good metric for an at-glance look at the functional significance of each clustering method we believe more analysis is necessary before we can fully argue for the benefits of Mapper. To do so we go further into our analysis of GDS1439 with two types of analysis. After choosing

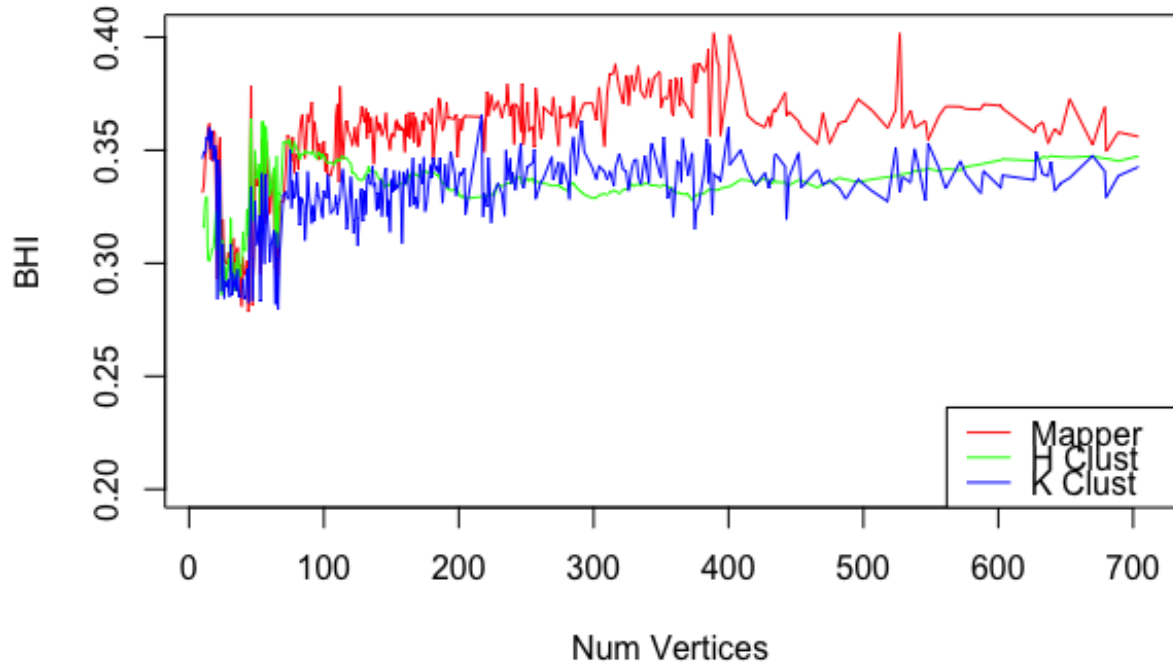


Figure 1: As the number of clusters rises, the BHI of Mapper clusterings separates itself

Table 1: BHI's for GDS1439

	Mapper	H-Clust	K-Means	Mapper-DEG	H-Clust-DEG	K-Means-DEG	Obs
Pearson	.3429 \pm .030	.3275 \pm .022	.3240 \pm .025	.3516 \pm .016	.3370 \pm .010	.3300 \pm .018	759
Spearman	.3538 \pm .013	.3462 \pm .012	.3396 \pm .038	.3139 \pm .039	.3285 \pm .032	.3400 \pm .038	520

Table 2: BHI's for GDS5437

	Mapper	H-Clust	K-Means	Mapper-DEG	H-Clust-DEG	K-Means-DEG	Obs
Pearson	.3074 \pm .012	.3266 \pm .017	.2990 \pm .006	.3023 \pm .022	.3205 \pm .022	.3218 \pm .033	264
Spearman	.3253 \pm .009	.3368 \pm .019	.2526 \pm .031	.2739 \pm .028	.2196 \pm .022	.2526 \pm .033	550

the parameters in which Mapper performed optimally (TODO WHICH), we compared the functions of the genes in each cluster, the pathways where those genes operate, as well as a Gold Standard analysis of the clustering destination of genes which are known to be related to prostate cancer.

3.2. Gene Function Analysis

Using the feature data in the GDS file we mapped each gene to its biological functions. For each biological function, we found the set of genes performing that function. We performed an enrichment analysis on the differentially expressed genes performing each gene function, using Fisher's Exact test. For each clustering method, we took the gene functions with the 15 smallest p-values and analyzed the clusters in which these genes were placed, //insert table here Genes performing the same gene function were clustered more closely by Mapper than by K-means or Hierarchical clustering. TODO what were the functions of these genes, whats the significance /TODO issue with GDS5437

3.3. Pathway Analysis

We used Reactome biological pathway database for our pathway analysis data. For each pathway, we found the set of genes acting in that pathway. We then performed a pathway enrichment analysis using Fisher's Exact test. We concentrated our analysis on the pathways with the 10 smallest p-values, and identified which genes act in these pathways and which clusters these genes end up in. //insert table The following pathways had their genes clustered in Mapper clusters 45,46,47 : Uptake of Oxygen and release of Carbon Dioxide by Erythrocytes, Uptake of Carbon Dioxide and release of Oxygen by Erythrocytes, Activation of genes by ATF4. Meaningful interactions between pathways can be discovered if the genes acting in those pathways are clustered together by Mapper. //has any research been done on these pathways //TODO issue with mice analysis

3.4. Gold Standard Analysis

To analyze the clustering we compiled a Gold Standard of genes of significance for GDS1439 from the Genetics Home Reference. On obtaining the optimum parameter combination of overlap, intervals and bins, for which the BHI values of Mapper are significantly higher than K-means and Hierarchical clustering, we analysed where the Gold Standard genes for each data set end up in each of the Clustering methods. Since our filter function for Mapper was the diseased component of data, and the Gold Standard for a disease is the set of genes which are expressed the most, Mapper should prove successful in clustering Gold Standard genes in similar clusters. An analysis of these clusters determines the effectiveness of Mapper. //insert table here On performing clustering analysis of the Gold Standard genes, we found that most of the genes end up between clusters 47-50, whereas no such pattern emerged in K-means or Hierarchical clustering. This proves that clustering using Mapper brings out biological significance between genes clustered together, when the right filter function is used. For GDS5437 we were not able to compile a Gold Standard of genes which are related to breast cancer, thus we were not able to perform a similar analysis.

4. Discussion

//nothing too specific to mapper, while things were good in 1439 they were not in 5437. Our discoveries on GDS1439 were promising. All of our methods of analysis showed Mapper to be a superior method in terms of functional analysis. On the other hand, our preliminary analysis on GDS5437 did not give any indication of an advantage to Mapper, and we were not able to perform any other sort of analysis on this data set.

Our goal, however, is to show that Mapper can be used as a method to discover functional significance, not that it will always produce clusterings of functional significance. One thing we noticed when working with Mapper initially was that it is very sensitive to parameter choices. Even in the introductory example of uncovering the underlying structure of a data set of points shaped like a figure 8, changing the given parameters slightly caused the algorithm to produce an output that seemingly had no relation to the initial data set. We believe that when working with this method it is necessary to spend significant amount of time discovering the optimal parameters. Only after these parameters have been found is it worth deeply analyzing the contents of the output.

It is also worth noting that regardless of the lack of gene function/pathway/gold standard analysis on GDS5437, the BHI results were not promising. There was no general trend that favored Mapper's clustering,

and Mapper’s clustering was only better than the other two with a frequency expected by chance. One possible reason for this was that the gene expression data was taken from lung tissue sample for mice with breast cancer. It is possible that because the tissue samples were not taken from the same region as the cancer that the diseased component of the data points was less significant. When we saw these results and realized we would not be able to perform any of our further types on analyses on this data we focused our efforts on GDS1439. Taking more time to carefully select a second dataset through which we would have been able to perform the full analysis is where we could have improved this research most.

5. Conclusion

If we were to go deeper into this analysis we would analyze more data sets to see if we could produce similar results. Only two data sets were used in this study and only one was successful. While we are arguing for the general abilities of this method, a positive result on one dataset is not enough to confirm our hypothesis. We believe our method of analysis provides a good framework for extending this research, and can be applied to other datasets with limited effort. In addition, there is nothing specific to cancer about our analysis. We believe this method is as transferrable to a wide range of diseases/disorders as it is to other cancer gene expression data.

6. 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