MA5113 Assignment 2: Class/Biomarker Discovery

Ty Medina

Part 2: Results Discussion

This report summarizes the findings based on the results obtained from Part 1: Data Analysis Procedure, in which the R package "multiClust" was used to generate sample clusters, survival curves, and differentially expressed genes from a microarray dataset.

Microarray gene expression measurements were obtained from the study published as "A genomic predictor of response and survival following taxane-anthracycline chemotherapy for invasive breast cancer." (Hatzis et al. 2011), which details a distant relapse-free survival study in HER2 negative breast cancer across 508 patients, which were subset into discovery and validation cohorts. In the study, patients were divided into ER negative and positive subgroups, as well as classified as either chemoresistant or chemosensitive, and classified by residual cancer burden. The study set out to use microarray data to mirror these classifications, and observe if this data was predictive of survival outcome after the appropriate adjuvant chemotherapy for the ER subtype.

By nature, the application of multiClust to this data set approaches the question differently, the largest difference being that the multiClust application proceeded with the entire undivided dataset with no patient phenotype information considered. The study instead observed each subtype's and classification's microarray data separately, then compared. This essentially creates something of a supervised vs. unsupervised difference, which naturally would produce varying results. Nonetheless, the multiClust approach is useful to show whether or not a naive approach can detect the same groups and outcomes, or even novel subgroups.

Gene Rankings

The genes obtained from the multiClust ranking functions through CV guided ranking and polynomial ranking share approximately 25% probes, when comparing the two selected fixed gene number methods. Although it cannot be determined from these lists which probes were found to be expressed most differentially, the lists were uploaded to Panther to determine if particular gene subsets were over represented.

Clustering

Two cluster numbers were used in this analysis: 3, a value predetermined from similar analyses; and 8, the value determined by gap statistic calculation through multiClust. However, it is unclear whether 8 was a value produced meritoriously from multiClust, or if it was a default value produced after failing to achieve an optimal solution, as no mention is made of this in the multiClust documentation. While both produced survival curves with significant P-values, the benefit of 8 clusters is dubious, and will be discussed later.

KMeans and hierarchical clustering both performed equally well, however it is difficult to determine whether they produced similar results, as each assigns cluster numbers differently. Without graphically representing the data through PCA or somehow manually permuting matches between cluster numbers to optimize matching, concurrency between the two methods could not be determined.

The original study in Hatzis et al. initially describes two groups, ER- and ER+, each of which received the appropriate, yet differing, treatment for each subtype. It is hypothesized that this should have produced the largest divide between clusters. Beyond this, it is difficult to say whether more clusters based on the cancer

phenotypes in the study would be expected from a naive approach. Heatmap comparison could possibly be used to further identify clustering patterns, however this too proved difficult.

Heatmap and Dendrograms

Unfortunately, the heatmaps and dendrograms produced for each clustering method in this study did not provide a clear view of any kind of cluster structure. Aside from obtaining a general view of which pathways were important for the clustering, differential expressions between groups were not evident.

Panther Analysis

Using the Panther database webtool, the genes from the three rankings were explored for some kind of structure. While a wide array of gene ontologies and functions were found, the most over-represented gene ontologies from the rankings were found to be blood circulation (Fold Enrichment=3.33, P-value=3.74E-3) and cell-matrix adhesion (Fold Enrichment=2.73, P-value=2.23E-4). As tumors are known to have altered cell matrix adhesion properties and vasculature, these ontologies may have implications in cancer type, particularly when considering metastasis and tumor angiogenesis.

In addition to ontologies, over-represented pathways were also explored, and included the Wnt signaling pathway, the Cadherin signaling pathway, and the chemokine/cytokine inflammation pathway, all three of which are implicated in breast cancer.

However, without evidence from the dendrograms and heatmaps, it cannot be said what role these might play in determining cancer subtype or outcome.

Survival

The original study attempted to discover differences in relapse-free survival based on cancer subtype and phenotype. In this analysis, it cannot be said which samples belonged to which subtypes. Indeed, it could be said that this is the point of attempting a naive clustering approach. However, without results detailing differential expression between clusters, any reasoning behind differential survival is impossible.

All 12 survival curves produced by multiClust using the selected methods did show significant P-values. Interestingly, using 3 clusters tends to show only 2 intuitively different survival outcomes, with 2 of the 3 clusters showing very similar outcome. This could possibly be attributed to the two major groups in the study: ER+, which showed higher survival, likely due to the possibility of endocrine therapy, and ER- which showed lower survival. While 8 clusters also showed significant P-value, the resolution between curves appears quite low, forming more of a distribution than discrete classifications. Because of this, it may not be particularly useful to

---

Conclusions

While significant differences were seen in survival between the clusters generated by multiClust, the lack of usable heatmap data prevents this method from being in any way predicative. Without understanding expression data per group, new data cannot be classified in any meaningful way. Furthermore, while pathways such as Wnt signaling appeared, the lack of expression data by cluster from the heatmap again makes this fairly unusable.