BDSI 2023

June 2023

1 Data

Integrative analyses of imaging and genomics data hold the potential to reveal biological insights into complex diseases like cancer. Our data was previously used for an integrative analysis of clinical outcomes in brain cancers, also known as radiogenomics. Briefly, radiogenomics finds associations between imaging outcomes obtained from radiological imaging modalities (e.g., MRI), with genomic markers. The significant associations between imaging and genomics are further used to model and predict clinical outcomes.

The data for this project has three components.

1. **Genomics**. The Genomic data is obtained from LinkedOmics (Vasaikar et al., 2017), a publicly available portal that includes data from multiple cancer types in The Cancer Genome Atlas (TCGA). This data contains 1289 gene pathway scores for 61 cancer patients.

   What are gene pathways and gene pathway scores? Broadly, a set of genes constitutes a gene pathway. The pathway membership of genes is derived from the Molecular Signature Database, a publicly available resource containing annotated genes divided into pathways.

   Normalized gene-level RNA sequencing measurements, produced by high-throughput sequencing, are converted into pathway scores. A pathway score assesses the relative variability of gene expression of genes in the pathway as compared to expression of genes not in the pathway. These scores are computed using the gene-set variation analysis (GSVA) procedure in Hänzelmann et al. [2013].

2. **Imaging**. The Imaging data is based on multi-institutional Magnetic Resonance Imaging (MRI) sequences available in The Cancer Imaging Archive (TCIA). This data contain a total of 143 imaging measurements for the same 61 cancer patients.

   What do the imaging measurements represent? Four types of MRI sequences are considered. Each of these sequences display different types of tissues with varying contrasts based on the tissue characteristics. First, voxel-level intensity values are measured from these sequences. However,

the raw intensity values are sensitive to the configuration of the MRI machine and are neither comparable across different subjects, nor between study visits for the same subject. To incorporate granular characteristics of tumor heterogeneity, a smoothed density is constructed from the voxel-level intensity histogram as done by Saha et al. [2016]. Finally, the variability in the intensity histograms, across different subjects, is captured through the scores from a principal component analysis on the space of density functions by using a Reimannian-geometric framework. These principal component scores, derived from MRI scans, are our imaging measurements.

3. Clinical Outcome. The clinical outcome in our data contains survival times for the same 61 patients (as the genomics and imaging data). This data has once again been collected from TCGA.

1.1 Exploratory Data Analysis Exercises

The following exercises are meant to familiarize you with the data in a way that, hopefully, will assist you in thinking about and coding your projects.

1.1.1 Survival Time (`Y`)

1. How many patients had survival times greater than 60?
2. What is the median survival time?
3. Does a boxplot of the survival times show any outliers?
4. Are there any missing survival times, i.e. are any coded as NA's?

1.1.2 Gene Pathway Scores (`pathway.scores`)

1. How many gene pathways are there in the dataset?
2. What is the most variation any gene pathway expresses across patients? What is the least? Variances or standard deviations are sufficient.
3. Expression of which pathway correlates most with expression of the `KEGG_DRUG_METABOLISM_CYTOCHROME_P450` pathway?
4. Among patients whose survival time was greater than 10, expression of which pathway is most correlated with survival time? (Hint: The elements of `Y` align with the rows of `pathway.scores`.)
5. Choose 10 gene pathways and produce a plot that displays their average expression for patients in the lower (0-25%) and upper (75-100%) quartiles of survival time.

1.1.3 Imaging Scores (`pc_scores`)

In this matrix, the column names note the MRI sequence (different sequences highlight different kinds of tissue), the subregion of the tumor, and the dimension of the principal subspace. For example, `T2_ED.3` is the projection of patients' T2 scans of the ED subregion of their tumor into the 3rd dimension, or principal component (PC), of the principal subspace.

1. How many PC's are included for each MRI/subregion combination?
2. Verify for the `T2_ED` MRI/subregion combination that the variation of the PC scores decreases in each additional dimension in the principal subspace (a vector of decreasing variances or standard deviations is sufficient).
3. Make a plot that shows for the first PC of each MRI/subregion combination the three gene pathways whose expressions correlate most and least with it and what those correlation values are.
4. Make a plot that shows the number of times each gene pathway is one of the three whose expressions most or least correlate with the first PC of an MRI/subregion combination.
5. (*Trickier) PCA finds the dimensions of a subspace in which variation of the data, when projected into that subspace, is greatest. Each PC score is a projection of the original data into a dimension of that subspace. For each MRI/subregion combination, what percentage of the variation in the projected data, i.e. the PC scores, can be explained by the first 2 PC's? (Hint: PC's are orthogonal to one another by construction, so variation in the direction of one PC is unique to that PC.)

References

Sonja Hänzelmann, Robert Castelo, and Justin Guinney. Gsva: gene set variation analysis for microarray and rna-seq data. *BMC bioinformatics*, 14:1–15, 2013.

Abhijoy Saha, Sayantan Banerjee, Sebastian Kurtek, Shivali Narang, Joonsang Lee, Ganesh Rao, Juan Martinez, Karthik Bharath, Arvind UK Rao, and Veerabhadran Baladandayuthapani. Demarcate: Density-based magnetic resonance image clustering for assessing tumor heterogeneity in cancer. *NeuroImage: Clinical*, 12:132–143, 2016.