The Use of Lattice Upstream Targeting for the Analysis of mRNA Expression for Cancers

LUST 2019

Tristan Holmes J.B. Nation et al

University of Hawaii at Manoa

tristanh314@gmail.com

PSU Systems Science Seminar
February 11, 2023

Presentation Overview

Abstract

In 2019 the UH Cancer Center hosted a project to identify genetic factors of interest in various types of cancer.

One of the results of this effort was the use of the Lattice Upstream Targeting (LUST) Algorithm to analyze mRNA expression data for 33 different types of cancer in the TCGA database. This effort will be the topic of this presentation.

The full results of this effort can be found on GitHub.

Results of a similar project conducted using data proprietary to the UH cancer center led to studies seeking to identfy new chemical treatments.

Overview of Procedure

- The LUST algorithm is a discrete mathematical method for analyzing continuous data, i.e., mRNA expression.
- For a given array of expression data, the algorithm is applied twice.
- 1. The first run is on the entire expression matrix and uses a graph theoretic objective function to rank the groups obtained. This pass identifies and ranks a small set of *metagenes* associated with the given cancer.
  2. The second run is on the expression matrix for each metagene and supervised by survival time as the objective function using the Fisher score to rank the results. This pass identifies small predictive *signiatures* for each metagene.
- In some cases, certain signiatures would seem approprite to use as guides for treatment.

Data Aquisition and Cleaning

- TCGA mRNA expression and clinical data are downloaded from the Broad Institute via the Firehose GDAC portal.
- Normalized gene expression files sequenced by Illumina HiSeq are used, reporting expression levels for 20,531 genes.
- Samples from tissue surrounding tumors are removed so that each patient has a single record representing tumor tissue.
- The expression data is log transformed, quantile normalized, and row centered.
- Survival times and censoring information for each patient are contained in the clinical data and used later in the process.

Data Discretization

- The expression data is represented by a $20531 \times N$ real valued matrix $\mathbf{E}$, where N is the number of samples.
- The matrix $\mathbf{E}$ is discretized into a $20531 \times N$ matrix $\mathbf{M}$ with entries in $\{-1, 0, 1\}$.
- The desired density D of non-zero entries in $\mathbf{M}$ is obtained by adjusting a threshold variable ϕ using the matrix secant method.
- For this study $D = 0.5$ for all cancers. In any particular study, one may seek to vary D to optimize the results.

Specifications

The LUST algorithm is used to find metagenes (*Part I*), or signiatures (*Part II*).

Input

- Discretized expression matrix **M**.
- Parameters *density* and *conftol*.
- For Part II only, clinical data such as survival.

Output

- Objective function ranked Metagenes (Part I) or signiatures (Part II).
- For Part II only, a score placing patients into high and low risk groups.
- For Part II only, Kaplan-Meyer survival curves.