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 12, 2023

Presentation Overview

1. Introduction

2. Data Setup

3. The Lattice Upstream Targeting Algorithm

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*, *conftol*, *overlap* and *noregs*.
- For Part II only, clinical data such as survival.

Output

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

Regulation and Equivalence

Assume the density D has been fixed (0.5 in this study). We use *conftol* (in this study 0.75 for Part I and either 0.66, 0.7, or 0.74 for Part II) to adjust sensitivity.

Definition

For a gene X, let X^+ denote the number of columns marked with 1 and x^- the number of columns marked with -1. We say X *regulates* Y, denoted $X \to Y$, if

1. $\frac{|X^+ \cap Y^+|}{|X^+|} \geq \mathit{conftol}, \text{ and}$
2. $\frac{|X^- \cap Y^-|}{|X^-|} \geq \mathit{conftol}.$

Definition

We say gene X is *eqvialent* to gene Y and write $X \approx Y$ if $X \to Y$ and $Y \to X$.

Forming Groups

The algorithm begins by computing, for each gene X

$$F_X := \{Y : Y \approx X\}$$

Note: F_X is not necessarily and equilence class as $\approx$ is not transitive. Different groups merged if

$$\frac{|F_X \cap F_Y|}{\min(|F_X|, |F_Y|)} \geq \mathit{overlap}$$

In this study, default values for *overlap* were 0.5 for Part I and 0.6 for Part II. Merging was performed only once.

Refinement Using Upstream Regulators

Score every gene X to measure it's effectiveness regulating the entire set of genes.

$$s_X = \frac{1}{N} \cdot \sum_{X \to Y} \frac{\left(|X^+ \cap Y^+| + |X^- \cap Y^-|\right)^2}{|X^+| + |X^-|}$$

Let G be a group from the previous step. For each $X \notin G$, conisder

$$G_X = \{X\} \cup \{Y \in G : X \to G\} ,$$

and assign a score

$$p_{X,G} = \frac{|G_X|}{|G|}(1 + s_X)$$

For $\mathit{noregs} = k$ (default 5), keep $G_{X_1}, \ldots, G_{X_k}$ with the k highest scores $p_{X,G}$ for further analysis.

Objective Functions

Part I

For a given metagene M with n genes, we consider M as a directed graph with edges determined by $X \to Y$, and let E be the set of edges of this graph.

We use a measure of the probability of obtaining a set of vertices of size n with $|E|$ edges.

$$f(M) = \frac{|E|}{n-1}$$

Objective Functions

Part II

For each G_X form a submatrix $\mathbf{E}_{G_X}$ from the undiscretized expression data.

Use eigen-survival analysis to produce a predictive score for each patient that is a linear combinatin of their expression values for G_X.

The top and bottom quartiles of the predictive scores are identified and used to calculate Kaplan-Meier expected surival curves.

Use the logrank and Cox tests to measure the separation of these two curves. Each test produces a p-value (p_1 and p_2, respectively). The *Fisher score* combines these measures to rank how well the signiature separates the survival curves.

$$F(G_X) = -\ln(p_1) - In(p_2)$$

Fase Discovery Rates - Notation

Fix a density D, let $p = \frac{D}{2}$, and let $\gamma = \mathit{conftol}$. Consider an $m \times n$ matrix with entries from $\{-1, 0, 1\}$ assigned from uniform probability distributions with densities $p,\ p,\ 1-2p$.

Probability row X has a entries 1 and b entries -1

$$g(n, a, b, p) = \binom{n}{a+b} p^{a+b}(1-2p)^{n-a-b}$$

Probability row Y has c entries 1 in a columns

$$h(a, c, p) = \binom{n}{c} p^{c}(1-p)^{n-c}$$

Fase Discovery Rates - Derivation

Fix a density D, let $p = \frac{D}{2}$, and let $\gamma = \mathit{conftol}$. Consider an $m \times n$ matrix with entries from $\{-1, 0, 1\}$ assigned from uniform probability distributions with densities $p,\ p,\ 1-2p$.

Probability $X \to Y$

$$\sum_{1 \le a,b \le n} g(n,a,b,p) \left(\sum_{a \ge c \ge \gamma a} h(a,c,p) \right) \left(\sum_{b \ge d \ge \gamma d} h(b,d,p) \right)$$

Expected number of relations $X \to Y$

$$E = m(m-1) \cdot \mathrm{prob}(X \to Y)$$

False Discovery Rates - In Practice

Testing on permuted data matrices shows these estimates are quite accurate for values of *conftol* used in this study.

The probability of random edges or used *conftol* is very low, on the order of 10^{-5} at most.

The worst-case scenario in this study was cholangiocarcinoma, with only 36 patients. Here E is about $9,220$, but the analysis found $830,000$ arrows.

For Part II, there are even fewer random arrows expected.

False Discovery Rates - In Practice

Testing on permuted data matrices shows these estimates are quite accurate for values of *conftol* used in this study.

The probability of random edges or used *conftol* is very low, on the order of 10^{-5} at most.

The worst-case scenario in this study was cholangiocarcinoma, with only 36 patients. Here E is about $9,220$, but the analysis found $830,000$ arrows.

For Part II, there are even fewer random arrows expected.

Sensitivity - Simulations

To test the sensitiivty of LUST, simulations were run $5,000 \times 120$ signal matrix $\mathbf{S}$ with a step signal in the first 200 rows consisting of 30 enttries of 1, 30 entries of -1, and 60 zeros.

A Gaussian noise matrix was made to create $\mathbf{M} = \mathbf{S} + a\mathbf{N}$, using a to adjust signal-to-noise ratio.

Repeated tests were run at various levels of *conftol*.

The conclusion was that the signals detected by Part I are quite strong.

Sensitivity - Results

Subtitle

SNR	Rows Found	False Positives
$-10db$	188	0
$-12.5db$	4	0
$-15db$	0	0

Table: *conftol* = 0.7

SNR	Rows Found	False Positives
$-10db$	200	0
$-12.5db$	196	0
$-15db$	50	0

Table: *conftol* = 0.6

Sensitivity - More Results

Subtitle

SNR	Rows Found	False Positives
$-10db$	200	0
$-12.5db$	200	0
$-15db$	199	4

Table: $conftol = 0.5$