The LUST Algorithm: A Discrete Mathematical Method for Analyzing Genetic Expression Data

LUST 2019

Tristan Holmes J.B. Nation et al

University of Hawaii at Manoa

tristanh314@gmail.com

PSU Systems Science Seminar
April 26, 2023

Presentation Overview

Abstract

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

Abstract

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

The result was the development of the Lattice Upstream Targeting (LUST) Algorithm to analyze mRNA expression data for 33 different types of cancer in the TCGA database.

Abstract

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

The result was the development of the Lattice Upstream Targeting (LUST) Algorithm to analyze mRNA expression data for 33 different types of cancer in the TCGA database.

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

Abstract

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

The result was the development of the Lattice Upstream Targeting (LUST) Algorithm to analyze mRNA expression data for 33 different types of cancer in the TCGA database.

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

Per J.B. Nation the UH Manoa Cancer Center separately used the results of these efforts to direct studies seeking to identfy new chemical treatments.

LUST at a Glance

- Input consists of **continuous** mRNA expression data for a set of patients whose tumors were biopsied, along with survival times from diagnosis.

LUST at a Glance

- Input consists of **continuous** mRNA expression data for a set of patients whose tumors were biopsied, along with survival times from diagnosis.
- The expression data is binned into “overexpression,” “underexpression,” and “medium expression.” A new array of **discrete** data is made with enries of -1, 1 and 0, respectively.

LUST at a Glance

- Input consists of **continuous** mRNA expression data for a set of patients whose tumors were biopsied, along with survival times from diagnosis.
- The expression data is binned into “overexpression,” “underexpression,” and “medium expression.” A new array of **discrete** data is made with enries of -1, 1 and 0, respectively.
- The **discrete** data is used to group genes that tend to overexpress and/or underexpress together. These groups are called *metagenes*.

LUST at a Glance

- Input consists of **continuous** mRNA expression data for a set of patients whose tumors were biopsied, along with survival times from diagnosis.
- The expression data is binned into “overexpression,” “underexpression,” and “medium expression.” A new array of **discrete** data is made with enries of -1, 1 and 0, respectively.
- The **discrete** data is used to group genes that tend to overexpress and/or underexpress together. These groups are called *metagenes*.
- The **discrete** expression data is analyzed again to find smaller sets of genes that regulate metagene expression, these smaller sets are called *signatures*.

LUST at a Glance

- Input consists of **continuous** mRNA expression data for a set of patients whose tumors were biopsied, along with survival times from diagnosis.
- The expression data is binned into “overexpression,” “underexpression,” and “medium expression.” A new array of **discrete** data is made with enries of $-1,\ 1$ and 0, respectively.
- The **discrete** data is used to group genes that tend to overexpress and/or underexpress together. These groups are called *metagenes*.
- The **discrete** expression data is analyzed again to find smaller sets of genes that regulate metagene expression, these smaller sets are called *signatures*.
- A regression θ is produced that takes **continuous** expression data for signatures as input and outputs a real number. Values of θ can be used to place patients in high and low risk categories.

Data Aquisition and Cleaning

- TCGA mRNA expression and clinical data are downloaded from the Broad Institute via the Firehose GDAC portal.

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.

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 the tissue surrounding the tumors are removed so that each patient has a single record representing tumor tissue.

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 the tissue surrounding the 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.

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 the tissue surrounding the 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.

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 $\{-1, 0, 1\}$.

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.

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*).

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.

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, Kaplan-Meyer survival curves and a regression model scoring each metagene based on survival.

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.

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 set of columns marked with 1 and X^- the set of columns marked with -1. We say X *regulates* Y, denoted $X \to Y$, if

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

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 set of columns marked with 1 and X^- the set of columns marked with -1. We say X *regulates* Y, denoted $X \to Y$, if

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

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

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

Forming Groups

The algorithm begins by computing, for each gene X

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

Note: F_X is not necessarily an equivalence class as $\approx$ is not transitive. Different groups are merged if

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

Forming Groups

The algorithm begins by computing, for each gene X

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

Note: F_X is not necessarily an equivalence class as $\approx$ is not transitive. Different groups are 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. For Part I, the resulting groups are then examined by hand to identify representative *metagenes*.

Objective Functions

Part I

For a given group from the previous step G with n genes, we consider M as a directed graph with edges determined by $X \rightarrow Y$, and let E be the set of edges of this graph.

Objective Functions

Part I

For a given group from the previous step G 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.

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

Objective Functions

Part I

For a given group from the previous step G 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.

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

An example of output for Part I can be found in [lust-2019, page 12]

Objective Functions

Part I

For a given group from the previous step G 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.

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

An example of output for Part I can be found in [lust-2019, page 12]

Choose representative G's for each clustering of groups, these representatives are the *metagenes* we analyze in Part II.

Refinement Using Upstream Regulators

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

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

Refinement Using Upstream Regulators

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

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

Let G be a representative group that was kept from the initial grouping step. For each $X \notin G$, conisder

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

Refinement Using Upstream Regulators

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

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

Let G be a representative group that was kept from the initial grouping step. For each $X \notin G$, conisder

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

and assign a score

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

Refinement Using Upstream Regulators

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

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

Let G be a representative group that was kept from the initial grouping 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 $noregs = k$ (default 5), keep $G_{X_1}, \ldots, G_{X_k}$ with the k highest scores $p_{X,G}$ for further analysis.

Eigen-Survival Analysis

For a signiature obtained from the previous step, form an expression matrix **M**. Consider the SVD:

$$\mathbf{M} = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^T .$$

Eigen-Survival Analysis

For a signiature obtained from the previous step, form an expression matrix **M**. Consider the SVD:

$$\mathbf{M} = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^T.$$

$$B = \left\{ j \mid \mathbf{v}_j \text{ significant in KM and Cox with } p \leq 0.5 \right\}$$

Eigen-Survival Analysis

For a signiature obtained from the previous step, form an expression matrix **M**. Consider the SVD:

$$\mathbf{M} = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^T.$$

$$B = \{j \mid \mathbf{v}_j \text{ significant in KM and Cox with } p \leq 0.5\}$$

$$\mathbf{w} = \sum_{j \in B} \text{sign}(j) \sigma_j \mathbf{v}_j$$

Eigen-Survival Analysis

For a signiature obtained from the previous step, form an expression matrix **M**. Consider the SVD:

$$\mathbf{M} = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^T.$$

$$B = \{j \mid \mathbf{v}_j \text{ significant in KM and Cox with } p \leq 0.5\}$$

$$\mathbf{w} = \sum_{j \in B} \text{sign}(j) \sigma_j \mathbf{v}_j$$

$$\mathbf{w} = \sum_{j \in B} \text{sign}(j) \mathbf{M}^T \mathbf{v}_j$$

Objective Functions

Part II

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

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.

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.

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. The *Fisher score* combines the p-values to rank the signiature.

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

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. The *Fisher score* combines the p-values to rank the signiature.

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

For metagenes with high Fisher scores, the eigen-survival score is a regression model that may be useful for classifying risk, see [melanoma]

False Discovery Rates - In Practice

The predicted number of random arrows is quite low.

False Discovery Rates - In Practice

The predicted number of random arrows is quite low.

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

False Discovery Rates - In Practice

The predicted number of random arrows is quite low.

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

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

False Discovery Rates - In Practice

The predicted number of random arrows is quite low.

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

The probability of random edges for used values of *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.

False Discovery Rates - In Practice

The predicted number of random arrows is quite low.

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

The probability of random edges for used values of *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 on a $5,000 \times 120$ 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.

Sensitivity - Simulations

To test the sensitiivty of LUST, simulations were run on a $5,000 \times 120$ 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.

Sensitivity - Simulations

To test the sensitiivty of LUST, simulations were run on a $5,000 \times 120$ 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*.

Sensitivity - Simulations

To test the sensitiivty of LUST, simulations were run on a $5,000 \times 120$ 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

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

Table: *conftol* = 0.7

Sensitivity - Results

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

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

Table: *conftol* = 0.5

Conclusions

- Several metagenes appear to be of interest across multiple types of tumors with several variations. Other metagenes are prominnet for only a single kind of tumor.

Conclusions

- Several metagenes appear to be of interest across multiple types of tumors with several variations. Other metagenes are prominnet for only a single kind of tumor.
- Metagenes with signiatures that result in the separation of Kaplan-Meier survival curves indicate biological processes of interest.

Conclusions

- Several metagenes appear to be of interest across multiple types of tumors with several variations. Other metagenes are prominnet for only a single kind of tumor.
- Metagenes with signiatures that result in the separation of Kaplan-Meier survival curves indicate biological processes of interest.
- Separating tumors by stage results in different metagenes of interest, seeming to indicate that different biological processes become more prominent as the disease progresses.

Future Investigations

- Using signiatures to determine a patient's risk and aggresiveness of treatment (Nation, 2019). Training is promising, further open source testing is needed.

Future Investigations

- Using signiatures to determine a patient's risk and aggresiveness of treatment (Nation, 2019). Training is promising, further open source testing is needed.
- Include methyliation and microRNA expression in the analysis.

Future Investigations

- Using signiatures to determine a patient's risk and aggresiveness of treatment (Nation, 2019). Training is promising, further open source testing is needed.
- Include methyliation and microRNA expression in the analysis.
- Modify the $X \rightarrow Y$ relationship to include negative correlation.

Future Investigations

- Using signiatures to determine a patient's risk and aggresiveness of treatment (Nation, 2019). Training is promising, further open source testing is needed.
- Include methyliation and microRNA expression in the analysis.
- Modify the $X \rightarrow Y$ relationship to include negative correlation.
- Use the algorithm to study continuous data related to other diseases, specifically where the diseased tissue can be isolated and sampled.

The Last Word

"LUST is good...

Future Investigations

- Using signiatures to determine a patient's risk and aggresiveness of treatment (Nation, 2019). Training is promising, further open source testing is needed.
- Include methyliation and microRNA expression in the analysis.
- Modify the $X \rightarrow Y$ relationship to include negative correlation.
- Use the algorithm to study continuous data related to other diseases, specifically where the diseased tissue can be isolated and sampled.

The Last Word

"LUST is good...

...and so is the algorithm." - J.B. Nation

References

Adiricheva, Nation, et al (2015)
Measuring the Implications of the D-basis in Analysis of Data in Biomedical Studies
github

Nation, Okimoto, et al (2019)
A Comparative Analyisis of mRNA Expression for 33 Different Cancers, Part 1: The LUST Algorithm
github

Nation (2019)
A Genetic Signiature Predicting Survival and Metastasis for Melanoma Patients
github

Acknowledgements

University of Hawaii

- Professor Emeritus J.B. Nation
- Professor Emeritus Ralph Freese
- Professor Emeritus Bill Lampe

Portland State University

- Professor Wayne Wakeland
- Professor Martin Zwick

Moral Support and Inspiration

- Friends and Family
- Agnes Meyer Driscoll
- Charles Lutwidge Dodgson

Special Assistant

- Doctor Frankenstein

Thank you!