The *Mapper* Algorithm and its Application

Ruth Lang Fuentes

October 23, 2019

Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival

Monica Nicolau[a], Arnold J. Levine[b,1], and Gunnar Carlsson[a,c]

[a]Department of Mathematics, Stanford University, Stanford, CA 94305; [b]School of Natural Sciences, Institute for Advanced Study, Princeton, NJ 08540; and [c]Ayasdi, Inc., Palo Alto, CA 94301

Contributed by Arnold J. Levine, February 25, 2011 (sent for review July 23, 2010)

Content

1. Introduction
2. Preliminary Mathematical Tools
3. Application of PAD to Breast Cancer Microarray Data
4. Discussion

- Introduction of a method that extracts information from high-dimensional and very sparse microarray data by using *Mapper* (PAD)
- Visualization of the data via a graph has an advantage over clustering
- Method was applied to breast cancer transcriptional data
- Identification of a unique subgroup of ER+ breast cancers with 100 per cent overall survival and no metastasis

Mathematical tools

Progression Analysis of Disease (PAD):
Application of the *Mapper* algorithm to DSGA (Disease- Specific Genomic Analysis)- transformed data

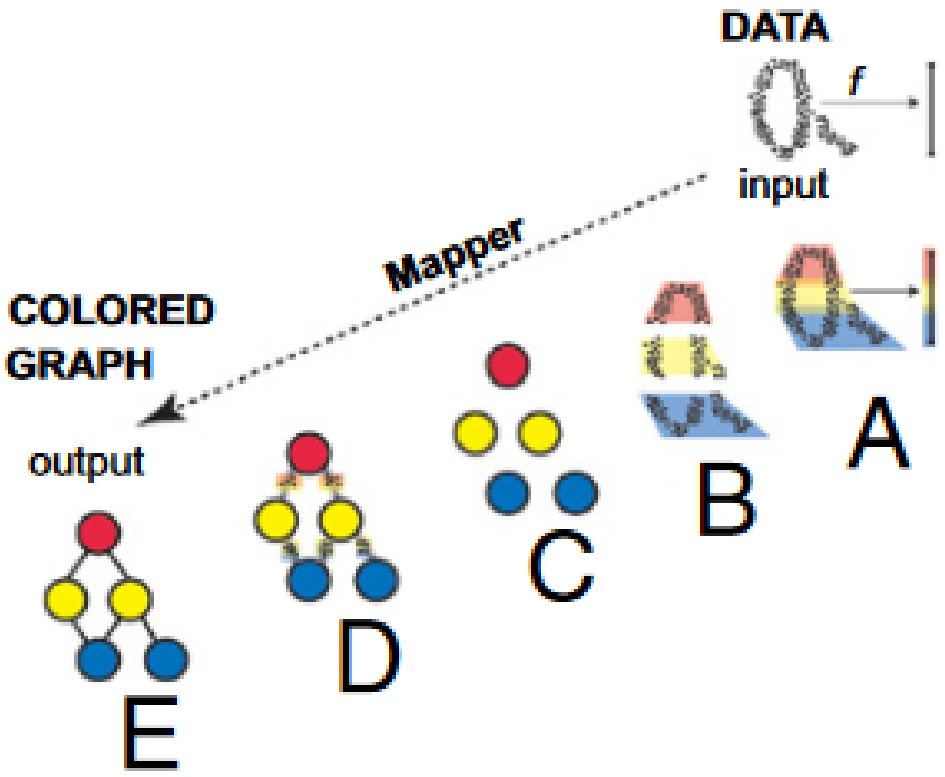
Recap: Mapper

KEY IDEA: Identify local clusters within the data and understand the interaction between them

- *Input:* Data set X + Filter function f + Distance notion
- *Output:* (Colored) graph

PRESERVES INFORMATION ABOUT ORIGINAL SHAPE, WHILE PROVIDING A SIMPLIFIED MATHEMATICAL OBJECT

Recap: Mapper



Disease-Specific Genomic Analysis (DSGA)

IDEA: Transform the data from diseased tissue as a sum of two terms :

- Normal component *Nc.* $\vec{T}$ mimics healthy tissue (obtained by computing a Healthy State Model)
- Disease component *Dc.* $\vec{T}$ measures error or deviation from normal and are the data used when applying *Mapper*

Let $\vec{T}$ be the original tumor vector, then:

$$\vec{T} = Nc.\vec{T} + Dc.\vec{T}.$$

DSGA HIGHLIGHTS THE DEGREE TO WHICH DISEASED TISSUE DATA ARE ABERRANT FROM HEALTHY TISSUE DATA

Disease-Specific Genomic Analysis (DSGA)

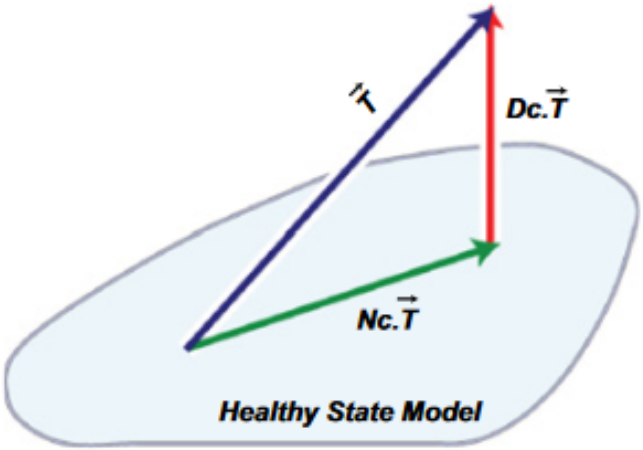


Fig. 2. *DSGA* decomposition of the original tumor vector $\vec{T}$ into the *Normal component* its linear models fit $Nc.\vec{T}$ onto the *Healthy State Model* and the *Disease component* $Dc.\vec{T}$ vector of residuals.

Progression Analysis of Disease (PAD)

Input: Data matrix from diseased tissue, in which columns are patients and rows are genomic variable type

1. DSGA-transform all of the data
2. Threshold data coordinates
3. Define *Mapper* filter functions for the columns

$$\vec{V} = <g_1, g_2, \ldots, g_s> .$$

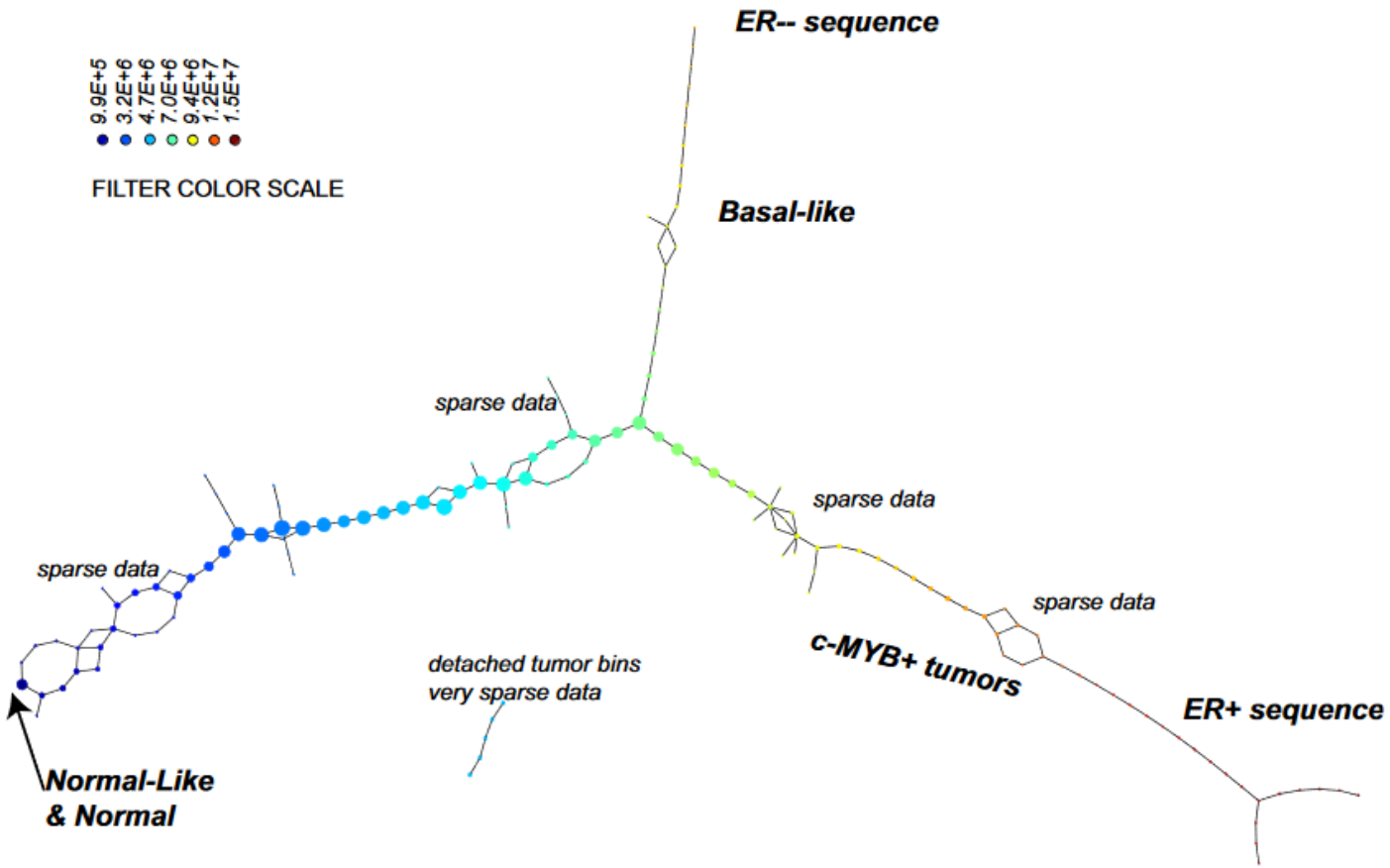
$$f_{p,k}(\vec{V}) = [\Sigma |\, g_r|^p]^{k/p} .$$

4. Apply *Mapper*, distance function on the data: Correlation distance.

Application of PAD to Breast Cancer Microarray Data

- Steps of PAD are now applied to a breast cancer microarray gene expression data set
- Step 1 and step 2 produced a data matrix with 262 rows (relevant genes)
- *Mapper* filter functions computed for p=1,...,5 and k=1,...,10
- Interpretation of the graph

PAD analysis for p=2 and k=4



c-MYB+ breast cancer

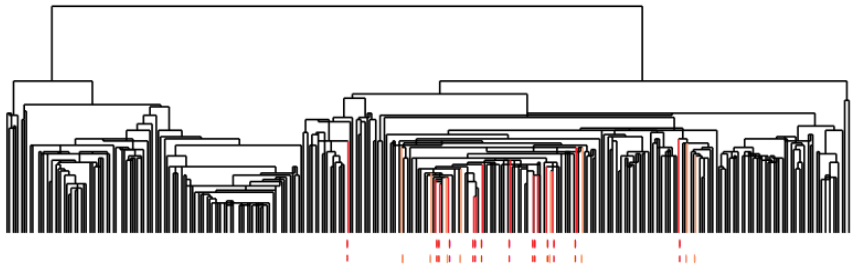
- 100 per cent overall survival with no recurrence and no death from disease (this information was **not** incorporated in the PAD analysis!)

- 100 per cent overall survival with no recurrence and no death from disease (this information was **not** incorporated in the PAD analysis!)
- This subgroup is not of one special molecular subtype (Molecular Subtype Classification)

- 100 per cent overall survival with no recurrence and no death from disease (this information was **not** incorporated in the PAD analysis!)
- This subgroup is not of one special molecular subtype (Molecular Subtype Classification)
- Prediction Analysis of Microarrays reveals that the subgroup is distinct from normal tissue and the *Normal-like* group and significantly homogeneous as a class (two predictor genes were able to distinguish between c-MYB+ group and normal tissue with error $= 0$)

- 100 per cent overall survival with no recurrence and no death from disease (this information was **not** incorporated in the PAD analysis!)
- This subgroup is not of one special molecular subtype (Molecular Subtype Classification)
- Prediction Analysis of Microarrays reveals that the subgroup is distinct from normal tissue and the *Normal-like* group and significantly homogeneous as a class (two predictor genes were able to distinguish between c-MYB+ group and normal tissue with error $= 0$)
- A set of genes that are significantly different between the c-MYB+ group and normal samples or the c-MYB+ group and the rest of the ER+ sequence was detected.

Comparison with Cluster Analysis



Clustering applied to the same DSGA-transformed data matrix.
The c-MYB+ tumors (in red) are scattered among different clusters.
Unlike PAD, cluster analysis was unable to identify this new group!

Conclusion:

c-MYB+ breast cancer is a unique group that does not fit into previously identified breast cancer types, shows uniformity in molecular signature and clinical and survival properties. It was also validated in other breast cancer data sets.

Discussion

- TDA is particularly appropiate for the analysis of biological data (combinatorial viewpoint, robustness to distortions)
- *Mapper* is more sensitive than cluster analysis in identifying the subtle geometry of (genomic) data

Thanks for listening!