Machine Learning Tools for Omics Data Analysis

Jing Ma

Public Health Sciences Division
Fred Hutch Cancer Research Center

29 October, 2020

Outline

Differential Network Enrichment Analysis

Generalized Principal Component Analysis

Part I: Differential Network Enrichment Analysis

Collaborators: Alla Karnovsky and Farsad Afshinnia from U of Michigan, George Michailidis from U of Florida, Ali Shojaie from U of Washington

Case Study: Chronic Kidney Disease

Motivation

- Two study cohorts:
  - Clinical Phenotyping Resource and Biobank Core (CPROBE[1])
  - Chronic Renal Insufficiency Cohort (CRIC[2])
- Lipids are important in different metabolic pathways along with diverse cellular and biological functions.

Goal

We want to identify lipidomic signatures of chronic kidney disease (CKD) progression. With two study cohorts, can assess replicability of results.

[1] Afshinnia et al. (2018) J. Am. Soc. Nephrol.

[2] Afshinnia et al. (2016) Kidney Int. Rep.

Lipidomics Data

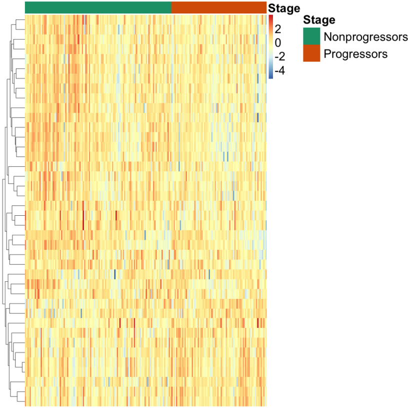


Fig. 1. Heatmap of DE lipids in CRIC

CPROBE

- early (n=135)
- advanced (n=79)

CRIC

- non-progressors (n=121)
- progressors (n=79)

Number of Features

- 285 lipids shared in both cohorts

Enrichment Analysis

Standard univariate approaches are easy to implement (e.g. t-test).

Enrichment Analysis

Standard univariate approaches are easy to implement (e.g. t-test).

But they

- suffer from low power,

Enrichment Analysis

Standard univariate approaches are easy to implement (e.g. t-test).

But they

- suffer from low power,
- can miss orchestrated changes,

Enrichment Analysis

Standard univariate approaches are easy to implement (e.g. t-test).

But they

- suffer from low power,
- can miss orchestrated changes,
- are difficult to interpret.

Enrichment Analysis

Standard univariate approaches are easy to implement (e.g. t-test).

But they

- suffer from low power,
- can miss orchestrated changes,
- are difficult to interpret.

Unfortunately, lipid pathways are not well characterized either.

Differential Network Enrichment Analysis

DNEA[3] is a data-driven approach for subnetwork discovery and enrichment analysis.

[3] Ma et al. (2019) Bioinformatics.

Differential Network Enrichment Analysis

DNEA[3] is a data-driven approach for subnetwork discovery and enrichment analysis.

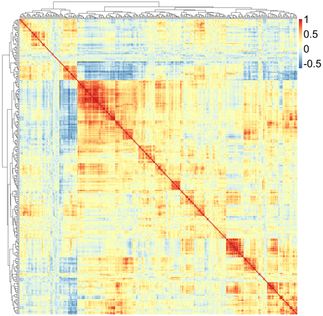
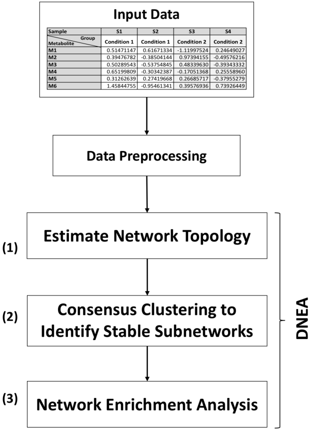


Fig. 2. Lipids are highly correlated!

[3] Ma et al. (2019) Bioinformatics.

Differential Network Enrichment Analysis



Input Data

Sample	S1	S2	S3	S4
Group Metabolite	Condition 1	Condition 1	Condition 2	Condition 2
M1	0.51471147	0.61671334	-1.11997524	0.24649027
M2	0.39476782	-0.38504144	0.97394155	-0.49576216
M3	0.50289543	-0.53754845	0.48339630	-0.39343332
M4	0.65199809	-0.30342387	-0.17051368	0.25558960
M5	0.31262639	0.27419668	0.26685717	-0.37955279
M6	1.45844755	-0.95461341	0.39576936	0.73926449

Data Preprocessing

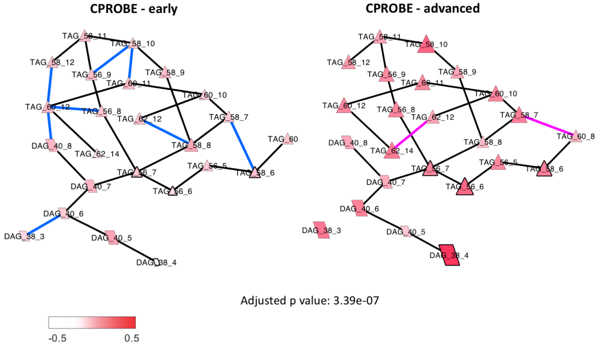
(1) Estimate Network Topology

(2) Consensus Clustering to Identify Stable Subnetworks

(3) Network Enrichment Analysis

DNEA

Novel Lipid Signatures in CKD



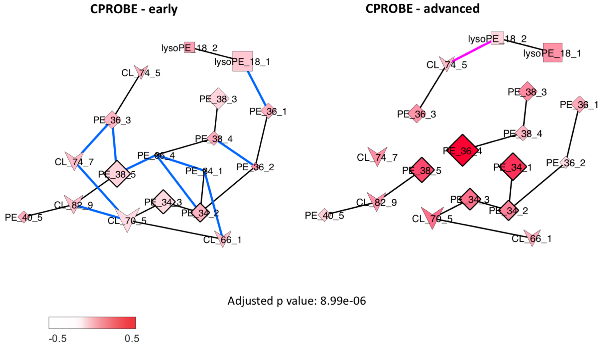
CPROBE - early

CPROBE - advanced

Adjusted p value: 3.39e-07

- Higher abundance of triacylglycerol (TAG) in advanced CKD
- Fewer edges were associated with advanced CKD

Novel Lipid Signatures in CKD



CPROBE - early

CPROBE - advanced

Adjusted p value: 8.99e-06

- Higher abundance of phosphatidylethanolamine (PE) in advanced CKD
- Fewer edges were associated with advanced CKD

DNEA Step I: Estimate Network Topology

What We Want

Two lipid co-expression networks for early and advanced stage, respectively.

DNEA Step I: Estimate Network Topology

What We Want

Two lipid co-expression networks for early and advanced stage, respectively.

☹ 135 and 79 observations in respective stage, but 285 lipids (features)

DNEA Step I: Estimate Network Topology

What We Want

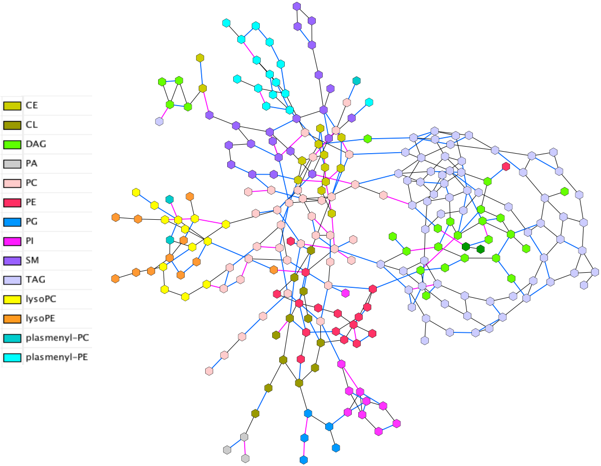
Two lipid co-expression networks for early and advanced stage, respectively.

☹ 135 and 79 observations in respective stage, but 285 lipids (features)

☺ Penalized estimation strategy to estimate sparse networks.

DNEA Step II: Identify Subnetworks

Apply consensus clustering to extract stable subnetworks (proxy to pathways)



DNEA Step III: Network Enrichment Analysis

Apply NetGSA[4] to detect enriched subnetworks.

[4] Ma et al. (2016) Bioinformatics; Ma et al. (2019) BMC Bioinformatics.

DNEA Step III: Network Enrichment Analysis

Apply NetGSA[4] to detect enriched subnetworks.

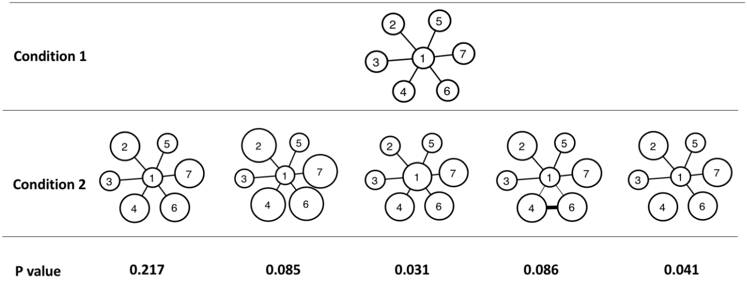


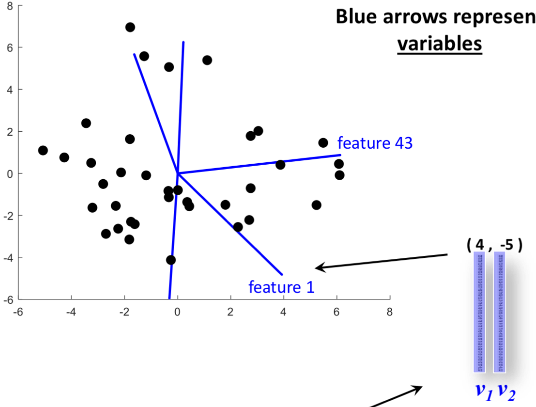
Fig. 3. NetGSA can detect changes in average expression and network topology

[4] Ma et al. (2016) Bioinformatics; Ma et al. (2019) BMC Bioinformatics.

Part II: Generalized Principal Component Analysis

Collaborators: Yue Wang from Arizona State, Tim Randolph from Fred Hutch,
Ali Shojaie from U of Washington

PCA for Visualization



Blue arrows represent <u>variables</u>

$$X \approx u_1 \, v_1^\top + u_2 \, v_2^\top$$

PCA for Visualization

☺ unsupervised

PCA for Visualization

☺ unsupervised

☹ does not allow prior knowledge

PCA for Visualization

☺ unsupervised

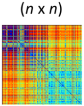
☹ does not allow prior knowledge

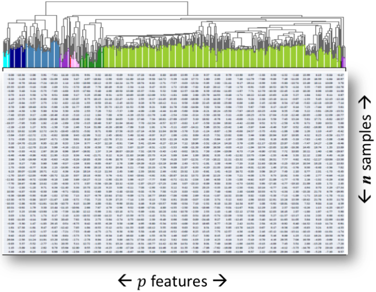
Examples of prior knowledge

- Topological relationships among features
- Non-Euclidean distances among observations

Our Framework: Generalized Matrix Decomposition

Allow prior information to be specified as kernel constraints[5].

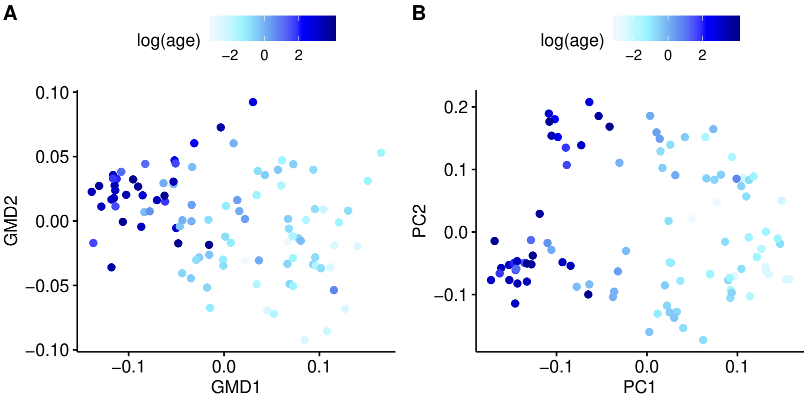




[5] Wang et al. (2019) mSystems

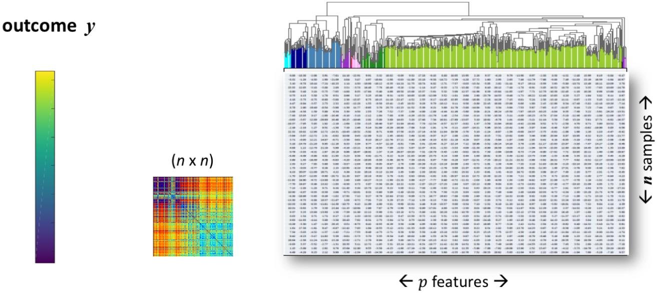
GMD vs PCA

Analysis of a microbiome dataset from Michelle et al. (2013) *Science*



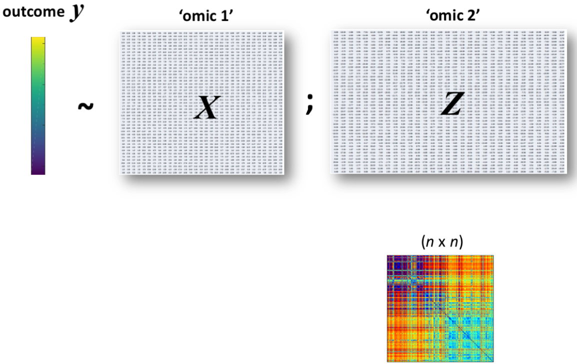
Supervised GMD

Identify features that are significantly associated with the outcome[6].



[6] Wang et al. (2020+)

GMD for Data Integration



Example: microbiome + metabolomics[7]; epigenetics + transcriptomics

[7] Wang et al. (2020+)

Summary

DNEA infers differential networks from experimental lipidomic or metabolomic data.

Summary

DNEA infers differential networks from experimental lipidomic or metabolomic data.

GMD allows dimension reduction and visualization, feature selection, and data integration.

Summary

DNEA infers differential networks from experimental lipidomic or metabolomic data.

GMD allows dimension reduction and visualization, feature selection, and data integration.

Big data? No problem!

References

1. Ma, Jing, et al. Differential network enrichment analysis reveals novel lipid pathways in chronic kidney disease. Bioinformatics (2019)
2. Ma, Jing, Ali Shojaie, and George Michailidis. Network-based pathway enrichment analysis with incomplete network information. Bioinformatics (2016)
3. Ma, Jing, Ali Shojaie, and George Michailidis. A comparative study of topology-based pathway enrichment analysis methods. BMC bioinformatics (2019).
4. Wang, Yue, et al. The generalized matrix decomposition biplot and its application to microbiome data. mSystems (2019)
5. Wang, Yue, et al. Generalized matrix decomposition regression: estimation and inference for two-way structured data. (2019).

drjingma.com