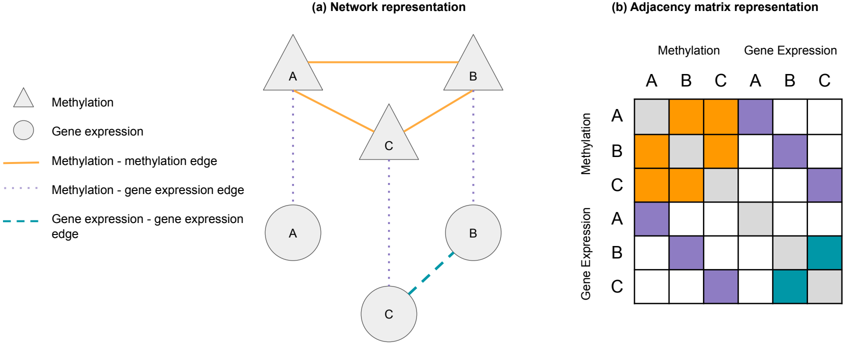
DRAGON: Determining Regulatory Associations using Graphical models on multi-Omic Networks

Kate Hoff Shutta

Des Weighill, Rebekka Burkholz, Marouen Ben Guebila, Helena U. Zacharias, Dawn L. DeMeo, John Quackenbush and Michael Altenbuchinger

NetBioMed NetSci Satellite 2022

Multi-omic networks



Gaussian Graphical Models (GGMs)

- In a GGM, we assume data are Gaussian and let edge weights represent partial correlations:

$$\rho_{X_i,X_j|V\setminus\{X_i,X_j\}} = \frac{Cov[X_i,X_j|X_{-ij}]}{\sqrt{Var[X_i|X_{-ij}]}\sqrt{Var[X_j|X_{-ij}]}} \tag{1}$$

- In this setting, absence of an edge $\iff$ zero partial correlation $\iff$ conditional independence

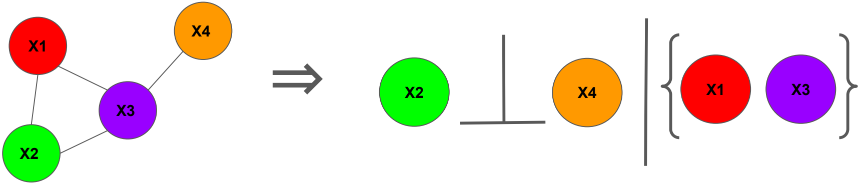


Figure 1: In this GGM, X_2 and X_4 are conditionally independent given X_1 and X_3.

Estimating a GGM from the precision matrix

- For $X_1, \dots, X_n \sim MVN(\mu, \Theta^{-1})$ with sample covariance matrix S, the precision matrix Θ can be converted to partial correlations, which are the edge weights of the GGM, using the relationship:

$$\rho_{X_i, X_j | V \setminus \{X_i, X_j\}} = -\frac{\Theta_{ij}}{\sqrt{\Theta_{ii}\Theta_{jj}}} \tag{2}$$

- Estimating a GGM is therefore equivalent to estimating Θ
- Easy case: invert the sample covariance: $\hat{\Theta} = S^{-1}$
- When is this a bad idea or impossible?
  - When there is high collinearity among the data, S^{-1} will be unstable
  - When $n < p$, S^{-1} does not exist

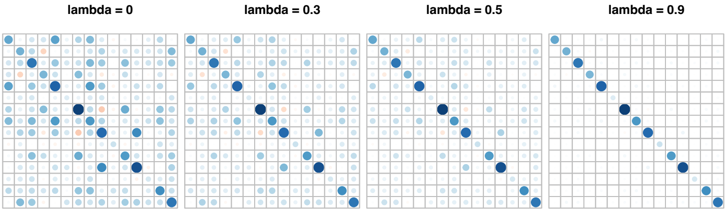
Covariance shrinkage for precision matrix estimation[1]

- For sample covariance S and target matrix T (e.g., $T = I_p$ and $T = diag(S)$), define

$$S' = (1-\lambda)S + \lambda T \tag{3}$$

- S' is not unbiased, but it is now invertible and can be used to estimate a precision matrix:

$$\hat{\Theta}_\lambda = (S')^{-1} = \{(1-\lambda)S + \lambda T\}^{-1} \tag{4}$$

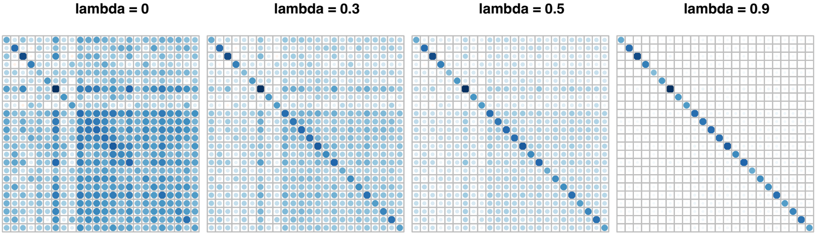


---

[1] Ledoit, O., Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. Journal of multivariate analysis, 88(2), 365-411.

Covariance shrinkage with two omics layers

- Toy example
  - $n = 15$ observations
  - $p_1 = 9$ predictors in omics layer 1
  - $p_2 = 15$ predictors in omics layer 2
  - Layer 2 has higher correlations
- Covariance shrinkage treats both layers identically, losing information from layer 1 because of the characteristics of layer 2



Extending Covariance Shrinkage with DRAGON

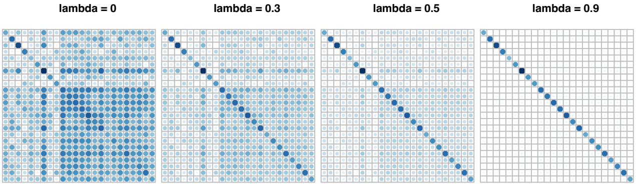
- DRAGON extends covariance shrinkage to handle different data types by introducing two shrinkage parameters:

$$\hat{\Theta} = \left[\begin{pmatrix} (1-\lambda_1) S^{(1,1)} & \sqrt{1-\lambda_1}\sqrt{1-\lambda_2} S^{(1,2)} \\ \sqrt{1-\lambda_1}\sqrt{1-\lambda_2} S^{(2,1)} & (1-\lambda_2) S^{(2,2)} \end{pmatrix} + \begin{pmatrix} \lambda_1 T^{(1)} & 0 \\ 0 & \lambda_2 T^{(2)} \end{pmatrix} \right]^{-1} \quad (5)$$

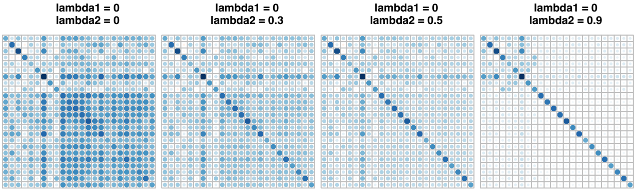
- Special cases of λ
  - Example 1: $\lambda_1 = 1$ or $\lambda_2 = 1$ "decouples" the technologies
  - Example 2: $\lambda_1 = \lambda_2$ yields $\hat{\Theta} = ((1-\lambda)S + \lambda T)^{-1}$; the estimator reduces to the original covariance shrinkage and the omics are treated the same

DRAGON enables omic-specific shrinkage

(a) Classical covariance shrinkage



(b) Covariance shrinkage with DRAGON

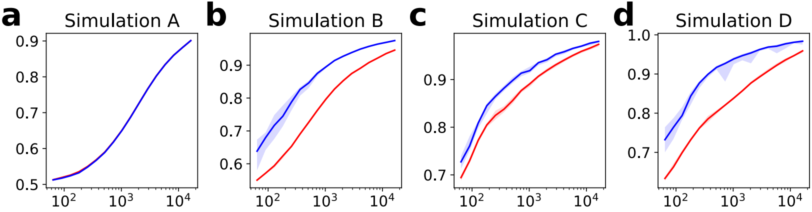


DRAGON outperforms standard GGMs in simulations

Table 1: Simulation study design; $p_1 = 100, p_2 = 500$ predictors in layer 1 and 2

Simulation	Edge Densities
A	$\eta_{(1,1)} = \eta_{(1,2)} = \eta_{(2,2)} = 0.05$
B	$\eta_{(1,1)} = \eta_{(1,2)} = 0.05; \eta_{(2,2)} = 0.005$
C	$\eta_{(1,1)} = 0.005; \eta_{(1,2)} = \eta_{(2,2)} = 0.05$
D	$\eta_{(1,1)} = 0.101, \eta_{(1,2)} = 0.01, \eta_{(2,2)} = 0.004$

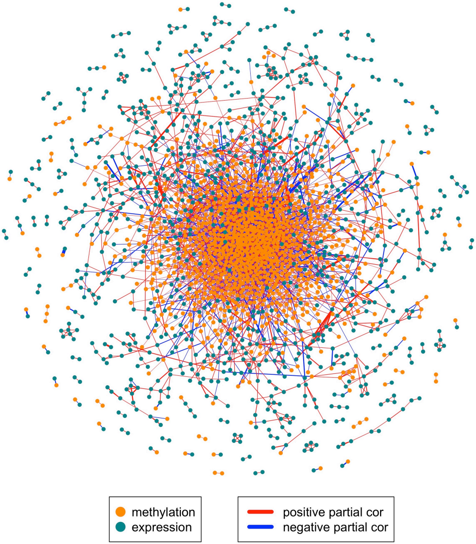
Figure 3: AUC for DRAGON (blue) and a standard GGM (red)



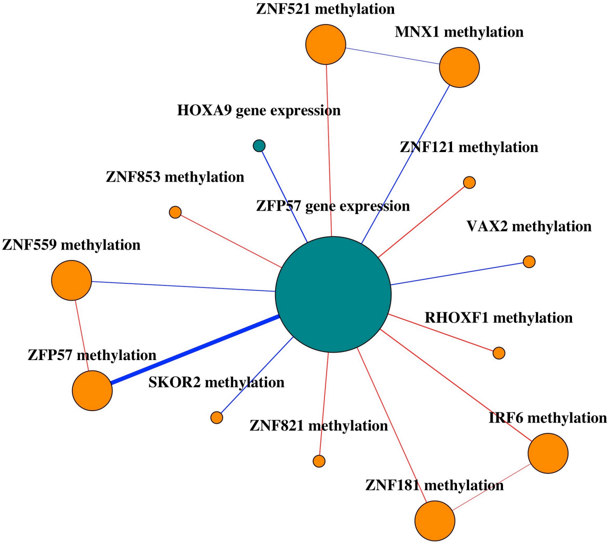
Application: multi-omic networks of promoter methylation and gene expression in breast cancer

- DNA methylation in the promoter region of a gene can affect the ability of the DNA to bind transcription factors and recruit the transcriptional machinery
- Typically, promoter methylation is associated with repression and demethylation is associated with activation of a gene
- To investigate these relationships in breast cancer, we built DRAGON networks on promoter methylation and gene expression of 1557 transcription factors based on samples from 765 women with breast cancer in The Cancer Genome Atlas (TCGA)

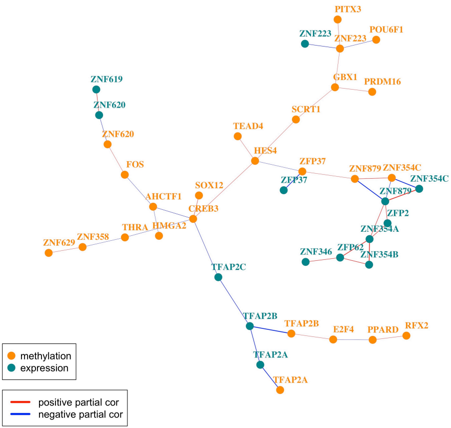
DRAGON Methylation-Gene Expression Network

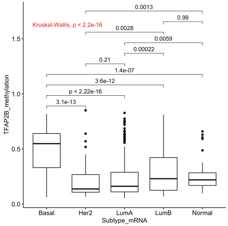


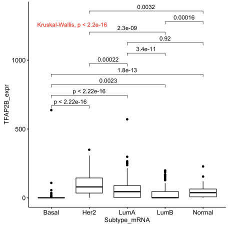
Multi-omic network hubs



Community detection for multi-omic biomarkers







Software and code availability

- DRAGON is available as Python software in the Network Zoo
  - Github: https://github.com/netZoo/netZooPy
  - Conda: https://anaconda.org/netzoo/netzoopy
- A Jupyter notebook showing how to use DRAGON is available at the NetworkZoo netbooks site: http://netbooks.networkmedicine.org/
- Code for the TCGA breast cancer analyses is available at https://github.com/katehoffshutta/DRAGON-TCGA-BRCA
- Methodology is described in our preprint: https://doi.org/10.48550/arXiv.2104.01690

Acknowledgments

- Des Weighill, Rebekka Burkholz, Marouen Ben Guebila, Helena U. Zacharias, Dawn L. DeMeo, John Quackenbush, Michael Altenbuchinger
- This work was supported by the NHLBI of the NIH under award number 2T32HL007427.
- The results shown here are in whole or part based upon data generated by the TCGA Research Network: `https://www.cancer.gov/tcga`. We gratefully acknowledge the participants of TCGA.

HARVARD
T.H. CHAN
SCHOOL OF PUBLIC HEALTH

NETWORK ZOO

Brigham and Women's Hospital
Founding Member, Mass General Brigham