Directed Cyclic Mixed Graph Modeling for Omic Data Integration

Carel F.W. Peeters
Dept. of Epidemiology & Biostatistics
VU University medical center
Amsterdam, The Netherlands
cf.peeters@vumc.nl

MRC Biostatistics Unit
Cambridge Institute of Public Health
Cambridge, UK
October 5, 2016

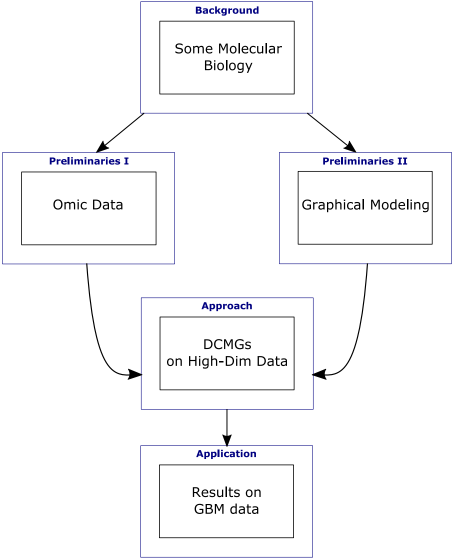
Contributors

Wessel N. van Wieringen
Dept. of Epimiology & Biostatistics, VUMC
Dept. of Mathematics, VU University Amsterdam

Anders E. Bilgrau
Novo Nordisk
Dept. of Mathematical Sciences, Aalborg University

Mark A. van de Wiel
Dept. of Epimiology & Biostatistics, VUMC
Dept. of Mathematics, VU University Amsterdam

Outline


Background
Some Molecular Biology
Preliminaries I
Omic Data
Preliminaries II
Graphical Modeling
Approach
DCMGs on High-Dim Data
Application
Results on GBM data

Omics and Omics Data

-ome

A totality of some (molecular biological) sort

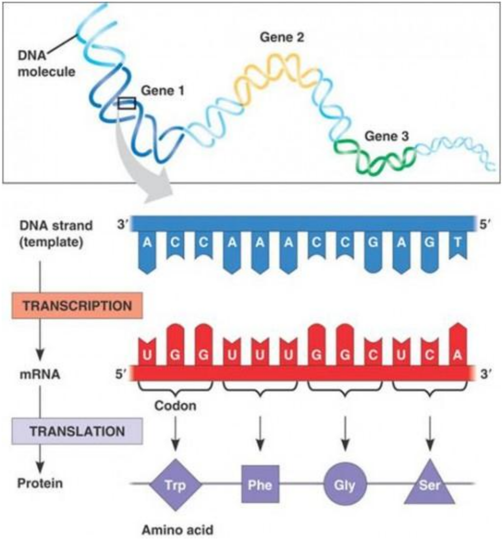
-omics

Collective quantification of some pool of molecular molecules

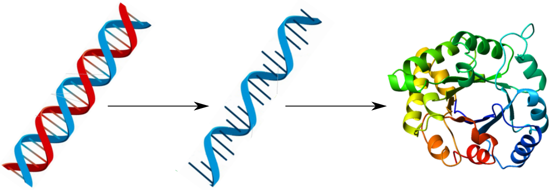
Genomics

The omics of the genome (of some organism)

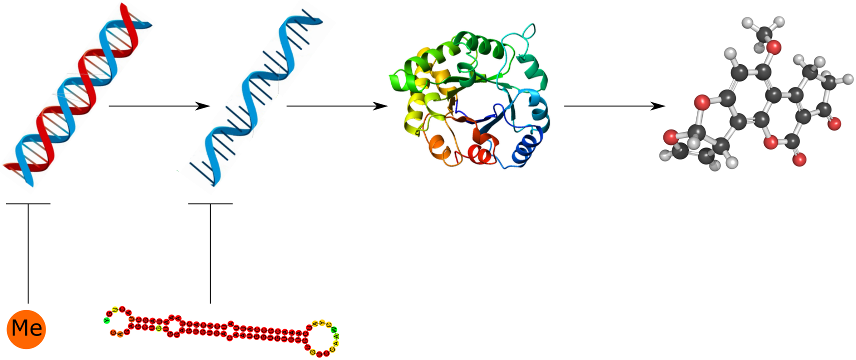
Central Dogma Molecular Biology



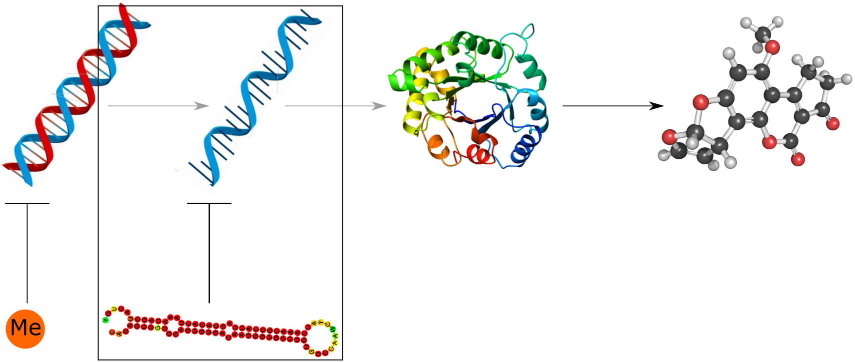
The Omic Cascade



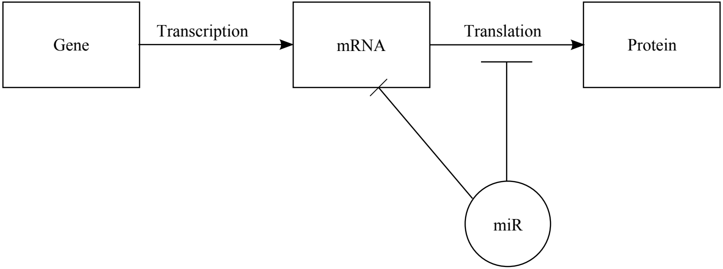
The Omic Cascade



The Omic Cascade



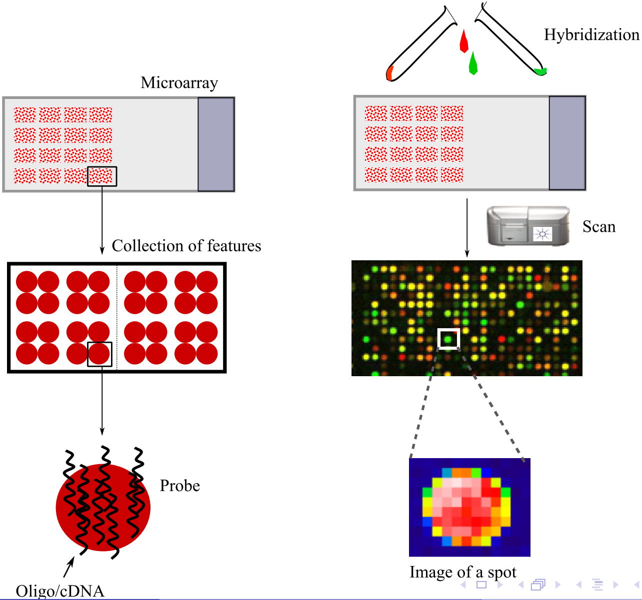
miRNA Epigenetics



micro RNA (miRNA)

- A family of small RNAs, approx. 22 nucleotides in length
- Bind to sequences of complementarity in target mRNA
- Post-transcriptional regulators of mRNA
- Logic: miRNA ↑ GE ↓; miRNA ↓ GE ↑
- RNA degradation or limiting of RNA translation
- Implicated in cancer

Array Data



Microarray
Collection of features
Probe
Oligo/cDNA
Hybridization
Scan
Image of a spot

Challenge: Dimensionality Omic Data

Variables

		1	2	3	⋯	p
Observations	1	y_{11}	y_{12}	y_{13}	⋯	y_{1p}
	2	y_{21}	y_{22}	y_{23}	⋯	y_{2p}
	3	y_{31}	y_{32}	y_{33}	⋯	y_{3p}
	4	y_{41}	y_{42}	y_{43}	⋯	y_{4p}
	5	y_{51}	y_{52}	y_{53}	⋯	y_{5p}
	⋮	⋮	⋮	⋮		⋮
	n	y_{n1}	y_{n2}	y_{n3}	⋯	y_{np}

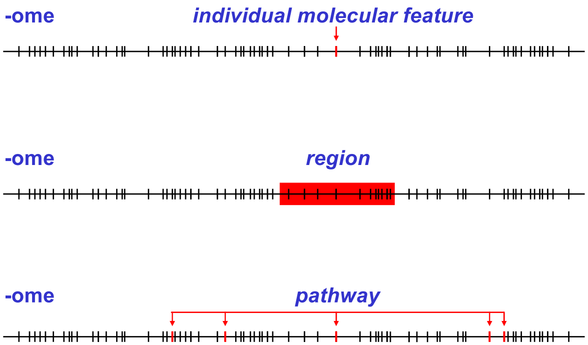
Molecular features

		1	2	3	4	5	⋯	p
Observations	1	y_{11}	y_{12}	y_{13}	y_{14}	y_{15}	⋯	y_{1p}
	2	y_{21}	y_{22}	y_{23}	y_{24}	y_{25}	⋯	y_{2p}
	3	y_{31}	y_{32}	y_{33}	y_{34}	y_{35}	⋯	y_{3p}
	⋮	⋮	⋮	⋮	⋮	⋮		⋮
	n	y_{n1}	y_{n2}	y_{n3}	y_{n4}	y_{n5}	⋯	y_{np}

Regular data: $n > p$

Omic data: $p > n$ or $p >> n$

Unit of Analysis



Gaussian Graphical Modeling

Graphical modeling

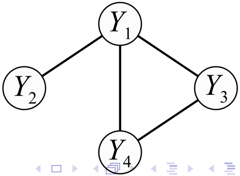
Class of models using graphs to express conditional (in)dependence relations between random variables

Gaussian setting

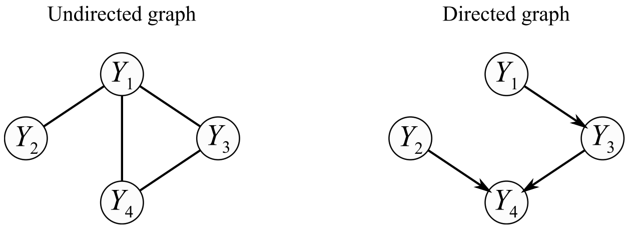
- Vertices: Correspond to random variables with normal distribution
- Edges: Correspond to the dependence structure
- Say $\mathbf{y} \sim \mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma})$, and define $\mathbf{\Sigma}^{-1} \equiv \mathbf{\Omega}$. Then, for $a, b \in$ vertex set V, $a \neq b$

$$-\frac{\omega_{ab}}{\sqrt{\omega_{aa}\omega_{bb}}} = 0 \Longleftrightarrow \omega_{ab} = 0 \Longleftrightarrow a \perp\!\!\!\perp b | V \setminus \{a, b\} \Longleftrightarrow a \not- b$$

$$\begin{bmatrix} \omega_{11} & \omega_{12} & \omega_{13} & \omega_{14} \\ \omega_{21} & \omega_{22} & 0 & 0 \\ \omega_{31} & 0 & \omega_{33} & \omega_{34} \\ \omega_{41} & 0 & \omega_{43} & \omega_{44} \end{bmatrix}$$



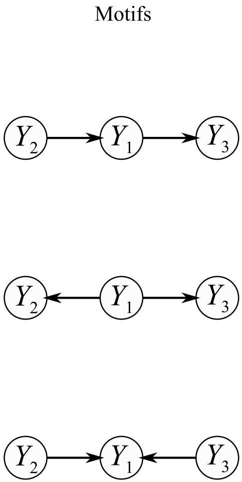
Undirected and Directed Graphs

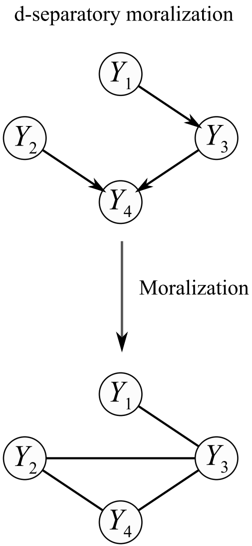


Directed Acyclic Graph (DAG)

$$\mathbf{y}_i := \mathbf{B}\mathbf{y}_i + \epsilon_i, \qquad i = 1, \ldots, n.$$

Directed Acyclic Graph (DAG)





Model and assumptions

Model

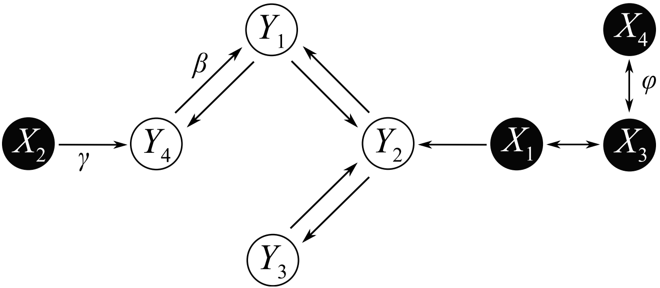
The SEM model we consider can be expressed as:

$$\mathbf{y}_i := \mathbf{B}\mathbf{y}_i + \boldsymbol{\Gamma}\mathbf{x}_i + \mathbf{I}_p\boldsymbol{\epsilon}_i, \qquad i = 1, \ldots, n.$$

Assumptions

1. Properly preprocessed data
2. $\mathbf{y}_i \perp\!\!\!\perp \mathbf{y}_{i'}, \forall i \neq i'$
3. $\boldsymbol{\epsilon}_i \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Psi})$, with $\boldsymbol{\Psi} \equiv \text{diag}[\psi_{11}, \ldots, \psi_{pp}]$, and $\psi_{jj} > 0, \forall j$
4. $\mathbf{x}_i \sim \mathcal{N}_q(\mathbf{0}, \boldsymbol{\Phi})$, with $\boldsymbol{\Phi} \succ 0$
5. $\mathbf{x}_i \perp\!\!\!\perp \boldsymbol{\epsilon}_{i'}, \forall i, i'$
6. $(\mathbf{I}_p - \mathbf{B})$ is nonsingular and $\beta_{jj} = 0, \forall j$

Graphical Representation: DCMG



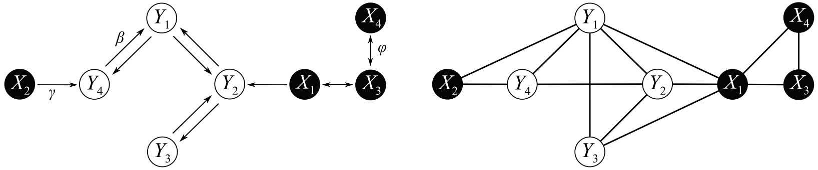
m-Separation

Stretching the idea of the collider

$$\longrightarrow \quad \longleftarrow$$

$$\longrightarrow \quad \longleftrightarrow$$

$$\longleftrightarrow \quad \longleftrightarrow$$



Some Results

Model-implied precision matrix

Let $\mathbf{y}_i := \mathbf{B}\mathbf{y}_i + \mathbf{\Gamma}\mathbf{x}_i + \mathbf{I}_p\epsilon_i$ be a SEM model satisfying assumptions 2-6. Define $\mathbf{\Theta} \equiv \{\mathbf{B}, \mathbf{\Gamma}, \mathbf{\Psi}, \mathbf{\Phi}\}$. Then $[\mathbf{y}_i^{\mathrm{T}}, \mathbf{x}_i^{\mathrm{T}}]^{\mathrm{T}} \sim \mathcal{N}_{(p+q)}[\mathbf{0}, \mathbf{\Sigma}(\mathbf{\Theta})]$, with

$$\mathbf{\Sigma}(\mathbf{\Theta})^{-1} \equiv \mathbf{\Omega}(\mathbf{\Theta}) = \begin{bmatrix} \mathbf{\Omega}(\mathbf{\Theta})_{yy} & \mathbf{\Omega}(\mathbf{\Theta})_{yx} \\ \mathbf{\Omega}(\mathbf{\Theta})_{xy} & \mathbf{\Omega}(\mathbf{\Theta})_{xx} \end{bmatrix} = \begin{bmatrix} (\mathbf{I}_p - \mathbf{B})^{\mathrm{T}}\mathbf{\Psi}^{-1}(\mathbf{I}_p - \mathbf{B}) & -(\mathbf{I}_p - \mathbf{B})^{\mathrm{T}}\mathbf{\Psi}^{-1}\mathbf{\Gamma} \\ -\mathbf{\Gamma}^{\mathrm{T}}\mathbf{\Psi}^{-1}(\mathbf{I}_p - \mathbf{B}) & \mathbf{\Phi}^{-1} + \mathbf{\Gamma}^{\mathrm{T}}\mathbf{\Psi}^{-1}\mathbf{\Gamma} \end{bmatrix}.$$

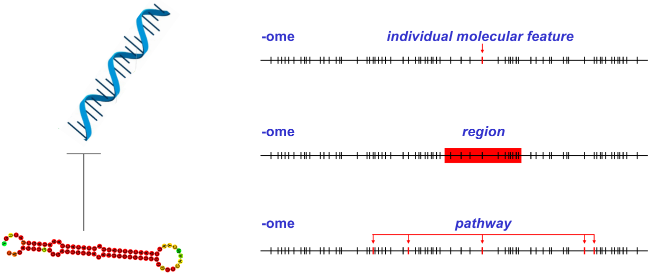
Identification by symmetric nonrecursion

If we assume that $\beta_{jk} = \beta_{kj}\ \forall j \neq k$, the model is (at least) locally identified

DCMG as graphical object

Assuming faithfulness, a perfect mapping can be shown

Step 1: Regularization



Setting

- Let $\hat{\mathbf{\Sigma}}$ denote the sample covariance matrix on $\mathbf{y}_i$ and $\mathbf{x}_i$
- When $(p+q) \rightarrow n$: $\hat{\mathbf{\Sigma}}$ is ill-behaved and $\hat{\mathbf{\Omega}} = \hat{\mathbf{\Sigma}}^{-1}$ is unstable
- When $(p+q) > n$: $\hat{\mathbf{\Sigma}}$ is singular and $\hat{\mathbf{\Omega}} = \hat{\mathbf{\Sigma}}^{-1}$ is undefined

Step 1: Regularization

Maximize

$$\underbrace{\ln|\mathbf{\Omega}| - \mathrm{tr}(\hat{\mathbf{\Sigma}}\mathbf{\Omega})}_{\text{log}-\text{likelihood}} - \underbrace{\frac{\lambda}{2}\|\mathbf{\Omega} - \mathbf{T}\|_2^2}_{\ell_2-\text{penalty}}$$

- $\mathbf{T}$ denotes a p.d. symmetric target matrix
- $\lambda \in (0, \infty)$ denotes a penalty parameter

Analytic penalized ML estimator

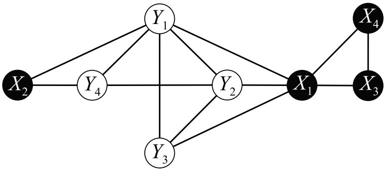
$$\hat{\mathbf{\Omega}}(\lambda) = \left\{ \left[\lambda \mathbf{I}_{(p+q)} + \frac{1}{4}(\hat{\mathbf{\Sigma}} - \lambda\mathbf{T})^2 \right]^{1/2} + \frac{1}{2}(\hat{\mathbf{\Sigma}} - \lambda\mathbf{T}) \right\}^{-1}$$

Step 2: Determine Support

Sparsified regularized precision

- Test for vanishing partial correlations to obtain $\hat{\mathbf{\Omega}}(\lambda)^0$
- A sparse representation of $\hat{\mathbf{\Omega}}(\lambda)$
- Local false discovery rate procedure

$$\left[\begin{array}{cccc|cccc}
\omega^{yy}_{11} & \omega^{yy}_{12} & \omega^{yy}_{13} & \omega^{yy}_{14} & \omega^{yx}_{11} & \omega^{yx}_{12} & 0 & 0 \\
\omega^{yy}_{21} & \omega^{yy}_{22} & \omega^{yy}_{23} & \omega^{yy}_{24} & \omega^{yx}_{21} & 0 & 0 & 0 \\
\omega^{yy}_{31} & \omega^{yy}_{32} & \omega^{yy}_{33} & 0 & \omega^{yx}_{31} & 0 & 0 & 0 \\
\omega^{yy}_{41} & \omega^{yy}_{42} & 0 & \omega^{yy}_{44} & 0 & \omega^{yx}_{42} & 0 & 0 \\
\hline
\omega^{xy}_{11} & \omega^{xy}_{12} & \omega^{xy}_{13} & 0 & \omega^{xx}_{11} & 0 & \omega^{xx}_{13} & \omega^{xx}_{14} \\
\omega^{xy}_{21} & 0 & 0 & \omega^{xy}_{24} & 0 & \omega^{xx}_{22} & 0 & 0 \\
0 & 0 & 0 & 0 & \omega^{xx}_{31} & 0 & \omega^{xx}_{33} & \omega^{xx}_{34} \\
0 & 0 & 0 & 0 & \omega^{xx}_{41} & 0 & \omega^{xx}_{43} & \omega^{xx}_{44}
\end{array}\right] \rightarrow$$



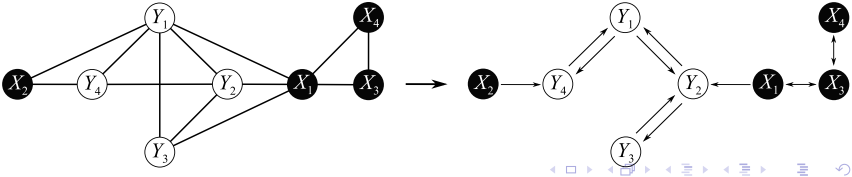
Step 3: Find DCMG

Parameter retrieval

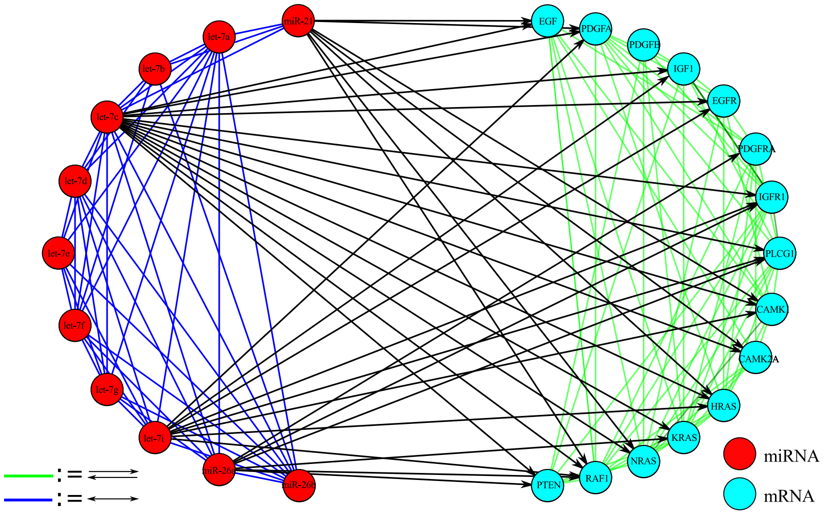
- From $\hat{\boldsymbol{\Omega}}(\lambda)^0$ we find $\hat{\boldsymbol{\Theta}}$ such that $\boldsymbol{\Omega}(\hat{\boldsymbol{\Theta}})$ is as close as possible to $\hat{\boldsymbol{\Omega}}(\lambda)^0$
- Inverse variance lemma and identification proposition imply simple iterative algorithm

Solving for Parameters

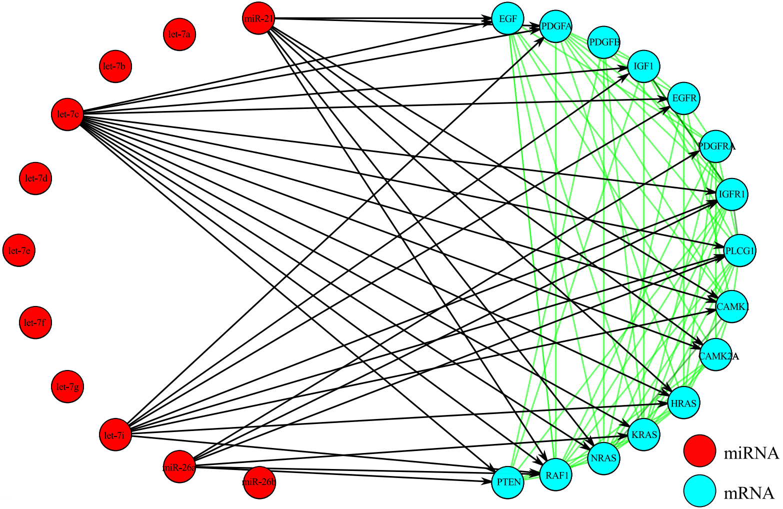
$$
\begin{aligned}
(\mathbf{I}_p - \mathbf{B}) &= \boldsymbol{\Psi}[\boldsymbol{\Psi}^{-1}\boldsymbol{\Omega}(\boldsymbol{\Theta})_{yy}]^{1/2} \\
\boldsymbol{\Gamma} &= -(\mathbf{I}_p - \mathbf{B})\boldsymbol{\Omega}(\boldsymbol{\Theta})_{yy}^{-1}\boldsymbol{\Omega}(\boldsymbol{\Theta})_{yx} \\
\boldsymbol{\Psi} &= [(\mathbf{I}_p - \mathbf{B})\boldsymbol{\Omega}(\boldsymbol{\Theta})_{yy}^{-1}(\mathbf{I}_p - \mathbf{B})] \circ \mathbf{I}_p
\end{aligned}
$$



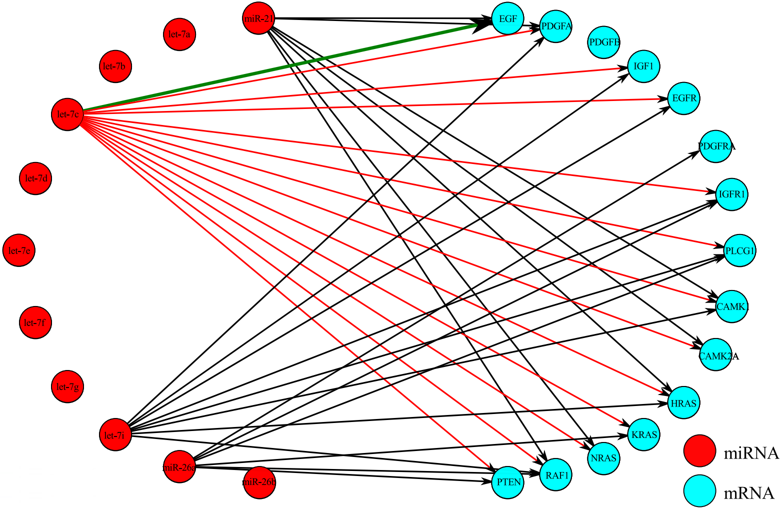
Full DCMG


miR-21
let-7a
let-7b
let-7c
let-7d
let-7e
let-7f
let-7g
let-7i
miR-26a
miR-26b
EGF
PDGFA
PDGFB
IGF1
EGFR
PDGFRA
IGFR1
PLCG1
CAMK1
CAMK2A
HRAS
KRAS
NRAS
RAF1
PTEN
miRNA
mRNA
:=
:=

Endogenous Relations



Exogenous Shocks



Koster, J.T.A. (1996) Markov Properties of Nonrecursive Causal Models. *Annals of Statistics*, 24:2148

Pearl, J. (2009, 2nd ed.) *Causality: Models, reasoning, and inference*. Cambridge, UK: Cambridge University Press

Richardson, T. (2003). Markov properties for acyclic directed mixed graphs. Scandinavian Journal of Statistics, 30: 145 157.

Schäfer, J., & K. Strimmer (2005) A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, 4:32

Software

- Peeters, C.F.W., Bilgrau, A.E., & van Wieringen, W.N. (2016). "rags2ridges: Ridge Estimation of Precision Matrices from High-Dimensional Data". R package, version 2.1.1 URL: https://cran.r-project.org/package=rags2ridges.

Theory/Methodology

- Peeters*, C.F.W., Bilgrau*, A.E., Eriksen, P.S., Boegsted, M., & van Wieringen, W.N. (2015). "Targeted Fused Ridge Estimation of Inverse Covariance Matrices from Multiple High-Dimensional Data Classes". arXiv:1509.07982v1 [stat.ME].
- Peeters, C.F.W., van Wieringen, W.N., & van de Wiel, M.A. (in preparation). "Directed Cyclic Mixed Graph Modeling for High-Dimensional Omic Data Integration".
- van Wieringen, W.N. & Peeters, C.F.W. (2016). "Ridge Estimation of Inverse Covariance Matrices from High-Dimensional Data". Computational Statistics & Data Analysis, 103: 284-303. arXiv:1403.0904v3 [stat.ME].

Computation

- Peeters, C.F.W., van de Wiel, M.A., & van Wieringen, W.N. (2016) "The Spectral Condition Number Plot for Regularization Parameter Determination". arXiv:1608.04123v1 [stat.CO].
- van Wieringen, W.N. & Peeters, C.F.W. (2015). "Application of a New Ridge Estimator of the Inverse Covariance Matrix to the Reconstruction of Gene-Gene Interaction Networks". In: di Serio, C., Lio, P., Nonis, A., and Tagliaferri, R. (Eds.) 'Computational Intelligence Methods for Bioinformatics and Biostatistics'. Lecture Notes in Computer Science, vol. 8623. Springer, pp. 170–179.

RAGS2RIDGES

2 Cents

"Get ridge or die trying"
- 2Cent