A multi-attribute Gaussian graphical model for inferring multiscale regulatory networks

An application in breast cancer

Julien Chiquet, MIA Paris

joint work with Martina Sundqvist, Guillem Rigaill
(original ideas with C. Ambroise, E. kolazcyk)

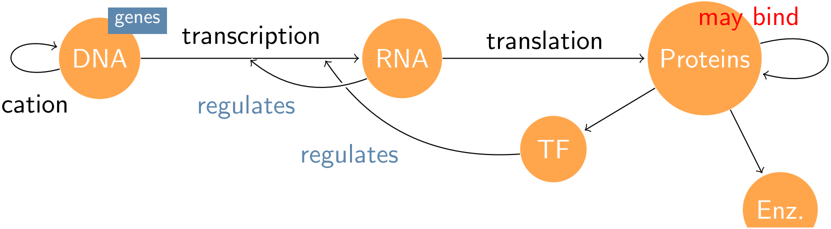
Statistiques aux Sommets, Rochebrune, 2018, March the 29th

J.C., G. Rigaill, M. Sundqvist,
Book on Gene Regulatory Networks: Methods and Protocols, Springer
Editors: Guido Sanguinetti, PhD and Vân Anh Huynh-Thu, PhD

multGGM package, development version on github

```
devtools::install_github("jchiquet/multGGM/multivarNetwork")
```

Why multi-attribute networks in genomics?



Data integration

- Omic technologies can profile cells at different levels: DNA, RNA, protein, chromosomal, and functional.
- multiple molecular profiles combined on the same set of biological samples can be *synergistic*.

Outline

1. Background on sparse GGM
2. Sparse multi-attribute GGM
3. Numerical experiments

Gaussian Graphical Model

Consider a size-p Gaussian random vector: $X \sim \mathcal{N}(\mathbf{0}_p, \mathbf{\Omega}^{-1})$.

- independence is equivalent to null covariance/correlation
- conditional independence is equivalent to null partial covariance/correlation

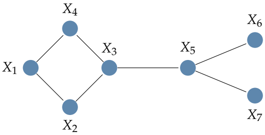
$$\rho_{ij} = -\mathbf{\Omega}_{ij} / \sqrt{\mathbf{\Omega}_{ii}\mathbf{\Omega}_{jj}}, \qquad \mathbf{\Omega}_{ii} = \mathbb{V}(X_i | X_{\backslash\{i,j\}})^{-1}$$

Conditional independence structure

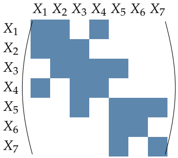
$$(i,j) \notin \mathcal{E} \Leftrightarrow Y_i \perp\!\!\!\perp Y_j | Y_{\backslash\{i,j\}} \Leftrightarrow \mathbf{\Omega}_{ij} = 0.$$

Graphical interpretation

$\mathcal{G} = (\mathcal{P}, \mathcal{E})$



$\mathbf{\Omega}$



	X_1	X_2	X_3	X_4	X_5	X_6	X_7
X_1	■	■		■			
X_2	■	■	■				
X_3		■	■	■	■		
X_4	■		■	■			
X_5			■		■	■	■
X_6					■	■	
X_7					■		■

⇝ Network reconstruction is (roughly) a variable selection problem

Gaussian Graphical Model and Linear Regression

Linear regression viewpoint

Gene expression X_j is linearly explained by the other genes':

$$X_j | X_{\setminus j} = - \sum_{k \neq j} \frac{\mathbf{\Omega}_{jk}}{\mathbf{\Omega}_{jj}} X_k + \varepsilon_j, \quad \varepsilon_j \sim \mathcal{N}(0, \mathbf{\Omega}_{jj}^{-1}), \quad \varepsilon_j \perp X_j$$

Conditional on its neighborhood, other profiles do not give additional insights

$$X_j | X_{\setminus j} = \sum_{k \in \mathsf{ne(j)}} \beta_{jk} X_k + \varepsilon_j \quad \text{with } \beta_{jk} = - \frac{\mathbf{\Omega}_{jk}}{\mathbf{\Omega}_{jj}}.$$

$\rightsquigarrow$ "Neighborhood" selection

Gold standard penalized approaches

Use ℓ_1 for both regularizing and promoting *sparsity*

Penalized likelihood (Banerjee *et al.*, Yuan and Lin, 2008)

$$\widehat{\mathbf{\Omega}}_{\lambda}^{\text{glasso}} = \arg\max_{\mathbf{\Omega}\in\mathcal{S}_p^+} \log\det(\mathbf{\Omega}) - \text{trace}(\mathbf{\Omega}\mathbf{S}_n) - \lambda \left\|\mathbf{\Omega}\right\|_{\ell_1}.$$

+ symmetric, positive-definite

– solved by the "Graphical-Lasso" ($\mathcal{O}(p^3)$, *Friedman et al, 2007*).

- R-packages **glasso**, **quic**, **huge**.

Neighborhood Selection (Meinshausen & Bülhman, 2006)

$$\hat{\mathbf{B}}^{\text{ns}} = \arg\min_{\mathbf{B}\in\mathbb{R}^{p\times p},\text{diag}(\mathbf{B})=\mathbf{0}_p} \frac{1}{2}\text{trace}(\mathbf{B}^\top\mathbf{S}_n\mathbf{B}) - \text{trace}(\mathbf{B}^\top\mathbf{S}_n) + \lambda\|\mathbf{B}\|_{\ell_1}.$$

– not symmetric, not positive-definite

+ p Lasso solved with Lars-like algorithms ($\mathcal{O}(npd)$ for d neighbors).

- R-package **huge**.

Gold standard penalized approaches

Use ℓ_1 for both regularizing and promoting *sparsity*

Penalized likelihood (Banerjee *et al.*, Yuan and Lin, 2008)

$$\widehat{\mathbf{\Omega}}_\lambda^{\text{glasso}} = \arg\max_{\mathbf{\Omega}\in\mathcal{S}_p^+} \log\det(\mathbf{\Omega}) - \operatorname{trace}(\mathbf{\Omega}\mathbf{S}_n) - \lambda \, \|\mathbf{\Omega}\|_{\ell_1}.$$

- + symmetric, positive-definite
- – solved by the “Graphical-Lasso” ($\mathcal{O}(p^3)$, *Friedman et al, 2007*).
- R-packages **glasso**, **quic**, **huge**.

Neighborhood Selection (Meinshausen & Bülhman, 2006)

$$\hat{\mathbf{B}}^{\text{ns}} = \arg\min_{\mathbf{B}\in\mathbb{R}^{p\times p},\text{diag}(\mathbf{B})=\mathbf{0}_p} \frac{1}{2}\operatorname{trace}\big(\mathbf{B}^\top\mathbf{S}_n\mathbf{B}\big) - \operatorname{trace}\big(\mathbf{B}^\top\mathbf{S}_n\big) + \lambda\|\mathbf{B}\|_{\ell_1}.$$

- – not symmetric, not positive-definite
- + p Lasso solved with Lars-like algorithms ($\mathcal{O}(npd)$ for d neighbors).
- R-package **huge**.

Gold standard penalized approaches (2)

Use ℓ_1 for both regularizing and promoting *sparsity*

CLIME – Pseudo-likelihood (Cai et al., 2011; Yuan, 2010)

$$\widehat{\mathbf{\Omega}}_{\lambda}^{\text{clime}} = \arg\min_{\mathbf{\Omega}} \left\|\mathbf{\Omega}\right\|_1 \text{ subjected to } \left\|\mathbf{S}_n\mathbf{\Omega} - \mathbf{I}\right\|_{\infty} \leq \lambda$$

– not positive-definite

+ p linear programs easily distributed ($\mathcal{O}(p^2 d)$ for d neighbors).

- R-package **fastclime** (dedictated imp. up to p=6!).

Sparse PArtial Correlation Estimation (SPACE) (Peng 2009; Khare 2014)

$$(\widehat{\boldsymbol{\rho}}_{\lambda}^{\text{space}}, \mathrm{diag}(\mathbf{\Omega})) = \arg\min_{\boldsymbol{\rho}, \mathrm{diag}(\mathbf{\Omega})} \frac{1}{2}\sum_{j=1}^{p} \omega_j \left\| \mathbf{x}_j - \sum_{k=1}^{p} \rho_{jk} \sqrt{\frac{\mathbf{\Omega}_{kk}}{\mathbf{\Omega}_{jj}}} \mathbf{x}_k \right\|_{\ell_2}^2 + \lambda \left\|\boldsymbol{\rho}\right\|_{\ell_1}$$

+ for fixed variances, same cost as neighborhood selection.

– alternate procedure without guarantees on the number of iterates

- R-package **gconcord**.

Gold standard penalized approaches (2)

Use ℓ_1 for both regularizing and promoting *sparsity*

CLIME – Pseudo-likelihood (Cai et al., 2011; Yuan, 2010)

$$\widehat{\mathbf{\Omega}}_\lambda^{\text{clime}} = \arg\min_{\mathbf{\Omega}} \|\mathbf{\Omega}\|_1 \text{ subjected to } \|\mathbf{S}_n\mathbf{\Omega} - \mathbf{I}\|_\infty \leq \lambda$$

– not positive-definite

+ p linear programs easily distributed ($\mathcal{O}(p^2 d)$ for d neighbors).

- R-package **fastclime** (dedictated imp. up to p=6!).

Sparse PArtial Correlation Estimation (SPACE) (Peng 2009; Khare 2014)

$$(\widehat{\boldsymbol{\rho}}_\lambda^{\text{space}}, \mathrm{diag}(\mathbf{\Omega})) = \arg\min_{\boldsymbol{\rho},\mathrm{diag}(\mathbf{\Omega})} \frac{1}{2}\sum_{j=1}^{p} \omega_j \left\| \mathbf{x}_j - \sum_{k=1}^{p} \rho_{jk}\sqrt{\frac{\mathbf{\Omega}_{kk}}{\mathbf{\Omega}_{jj}}}\mathbf{x}_k \right\|_{\ell_2}^2 + \lambda \|\boldsymbol{\rho}\|_{\ell_1}$$

+ for fixed variances, same cost as neighborhood selection.

– alternate procedure without guarantees on the number of iterates

- R-package **gconcord**.

Practical implications of theoretical results

Selection consistency (Ravikumar, Wainwright, 2009-2012)

Denote $d = \max_{j\in\mathcal{P}}(\mathrm{degree_j})$. Consistency for an appropriate λ and

- $n \approx \mathcal{O}(d^2 \log(p))$ for the graphical Lasso and Clime.
- $n \approx \mathcal{O}(d \log(p))$ for neighborhood selection (sharp).

(Irrepresentability) conditions are not strictly comparable...

Ultra high-dimension phenomenon (Verzelen, 2011)

Minimax risk for sparse regression with d-sparse models: useless when

$$\frac{d\log(p/d)}{n} \geq 1/2, \qquad (\text{e.g.}, n = 50, p = 200, d \geq 8).$$

Good news! when n is small, we don't need to solve huge problems because they can't but fail.

Model selection

Cross-validation

Optimal in terms of prediction, not in terms of selection

Information based criteria

- GGMSelect (Girault *et al*, '12) selects among a family of candidates.
- Adapt IC to sparse high dimensional problems, e.g.

$$\text{EBIC}_\gamma(\widehat{\mathbf{\Omega}}_\lambda) = -2\text{loglik}(\widehat{\mathbf{\Omega}}_\lambda; \mathbf{X}) + |\mathcal{E}_\lambda|(\log(n) + 4\gamma\log(p)).$$

Resampling/subsampling

Keep edges frequently selected on an range of λ after sub-samplings

- Stability Selection (Meinshausen and Bühlman, 2010, Bach 2008)
- Stability approach to Regularization Selection (StaRS) (Liu, 2010).

Outline

1. Background on sparse GGM

2. Sparse multi-attribute GGM

3. Numerical experiments

Multiattribute GGM

Consider e.g. some p genes of interest and the $K = 2$ omic experiments

1. X_{i1} is the expression profile of gene i (transcriptomic data),
2. X_{i2} is the corresponding protein concentration (proteomic data).

Define a block-wise precision matrix

- $X = (X_1, \dots, X_p)^T \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ in $\mathbb{R}^{pK}$,
- $X_i = (X_{i1}, \dots, X_{iK})^{\mathsf{T}} \in \mathbb{R}^K$.

$$\mathbf{\Omega} = \mathbf{\Sigma}^{-1} = \begin{bmatrix} \mathbf{\Omega}_{11} & & \mathbf{\Omega}_{1p} \\ & \ddots & \\ \mathbf{\Omega}_{p1} & & \mathbf{\Omega}_{pp} \end{bmatrix}, \qquad \mathbf{\Omega}_{ij} \in \mathcal{M}_{K,K},\ \forall (i,j) \in \mathcal{P}^2.$$

Graphical Interpretation

Define $\mathcal{G} = (\mathcal{P}, \mathcal{E})$ as the multivariate analogue of the *conditional graph*:

$$(i,j) \in \mathcal{E} \Leftrightarrow \mathbf{\Omega}_{ij} \neq \mathbf{0}_{KK}.$$

Multiattribute GGM as multivariate regression

Multivariate analysis view point

Straightforward algebra and we have

$$X_j \mid X_{\setminus j} = x \sim \mathcal{N}(-\boldsymbol{\Omega}_{jj}^{-1}\boldsymbol{\Omega}_{j\setminus j}x, \boldsymbol{\Omega}_{ii}^{-1}) \; .$$

or equivalently, letting $\mathbf{B}_j^T = -\boldsymbol{\Omega}_{jj}^{-1}\boldsymbol{\Omega}_{i\setminus j}$,

$$X_j \mid X_{\setminus j} = \mathbf{B}_j^T X_{\setminus j} + \boldsymbol{\varepsilon}_j \quad \boldsymbol{\varepsilon}_j \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Omega}_{ii}^{-1}), \quad \boldsymbol{\varepsilon}_j \perp X.$$

Remembering the univariate case?

$$X_j|X_{\setminus j} = -\sum_{k \in \text{neighbors}(j)} \frac{\boldsymbol{\Omega}_{jk}}{\boldsymbol{\Omega}_{jj}} X_j + \varepsilon_j, \quad \varepsilon_j \sim \mathcal{N}(0, \boldsymbol{\Omega}_{jj}^{-1}), \quad \varepsilon_j \perp X.$$

A matter of notation. . . I

Matrix of regression coefficients

$\mathbf{B}_j \in \mathcal{M}_{(p-1)K,K}$ is defined block-wise

$$\mathbf{B}_j = \left[\begin{array}{c} \mathbf{B}_j^{(1)} \\ \hline \vdots \\ \hline \mathbf{B}_j^{(j-1)} \\ \mathbf{B}_j^{(j+1)} \\ \vdots \\ \hline \mathbf{B}_j^{(p)} \end{array}\right] = - \left[\begin{array}{c} \mathbf{\Omega}_{j1} \\ \hline \vdots \\ \hline \mathbf{\Omega}_{j(j-1)} \\ \mathbf{\Omega}_{j(j+1)} \\ \vdots \\ \hline \mathbf{\Omega}_{j(p)} \end{array}\right]^\top \times \mathbf{\Omega}_{jj}^{-1},$$

$\rightsquigarrow$ the $K \times K$ matrix $\mathbf{B}_j^{(i)}$ links attributes of variables (i, j).

A matter of notation. . . II

Data matrix

Consider an i.i.d. sample $\{X^\ell\}_{\ell=1}^n$ of X such that each attribute is observed n times for the p variables

- $\mathbf{x}^\ell$ is a pK-size row vector
- $\mathbf{X}_j \in \mathcal{M}_{n,K}$ contains the data related to variable j
- $\mathbf{X}$ is the full data matrix in $\mathcal{M}_{n,pK}$

$$\mathbf{X} = \left[\begin{array}{c} \mathbf{x}^1 \\ \hline \vdots \\ \hline \mathbf{x}^n \end{array}\right] = \begin{bmatrix} \mathbf{X}_1 & \dots & \mathbf{X}_p \end{bmatrix}$$

$$= \left[\begin{array}{ccc|c|ccc} X_{11}^1 & \dots & X_{1K}^1 & \dots & X_{p1}^1 & \dots & X_{pK}^1 \\ \hline \vdots & & \vdots & \dots & & & \\ \hline X_{11}^n & \dots & X_{1K}^n & \dots & X_{p1}^n & \dots & X_{pK}^n \end{array}\right].$$

Multivariate neighborhood selection

The penalized multivariate regression approach

For each node /gene, recover its neighborhood by solving

$$\arg\min_{\mathbf{B}_j \in \mathcal{M}_{(p-1)K,K}} \frac{1}{2n} \left\| \mathbf{X}_j - \mathbf{X}_{\backslash j} \mathbf{B}_j \right\|_F^2 + \lambda \,\mathsf{Pen}(\mathbf{B}_j),$$

Choice of penalty

Group-based penalty to activate the set of attributes simultaneously on a given link:

$$\mathsf{Pen}(\mathbf{B}_j) = \sum_{k \neq j} \|\mathbf{B}_j^{(k)}\| \ , \quad \mathbf{B}_j^{(k)} \in \mathcal{M}_{KK}$$

- $\|M\| = \|M\|_F = \left(\sum_{i,j} M_{ij}^2\right)^{1/2}$, the Frobenius norm,
- $\|M\| = \|M\|_\infty = \max_{i,j} |M_{ij}|$, the sup norm (shared magnitude),
- $\|M\| = \|M\|_\star = \sum \mathrm{eig}(M)$, the nuclear norm (rank penalty).

Outline

1. Background on sparse GGM
2. Sparse multi-attribute GGM
3. Numerical experiments

Simulation study: settings

1. Draw a $p \times p$ adjacency matrix $\mathbf{A}$ under Erdös-Renyi model.
2. Expand $\mathbf{A}$ to multivariate space:

$$\mathbf{M} = \mathbf{A} \otimes \mathbb{S} + \mathbf{I}_{p \times K}$$

   $\mathbb{S}$ is used to consider different scenarios of agreement

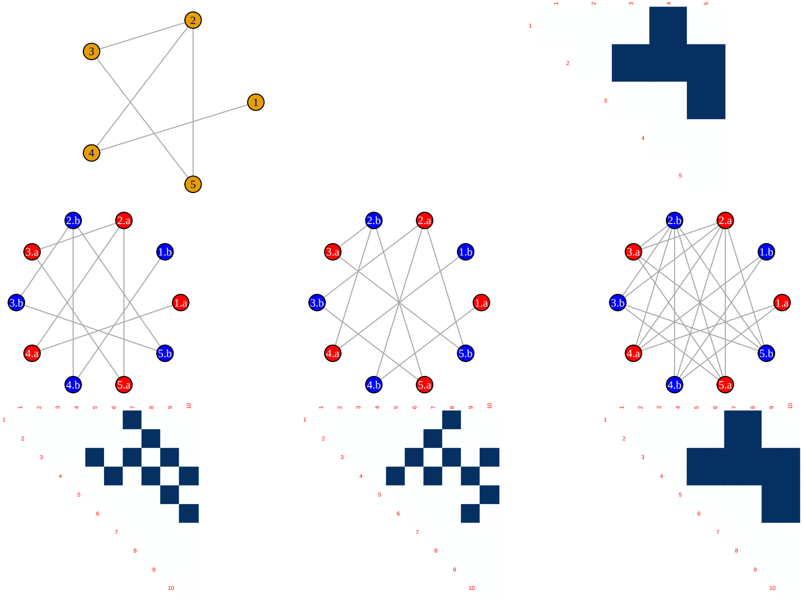
   a) $\mathbb{S} = \mathbf{I}_{K,K}$
      $\leadsto$ same intra-attribute network, no inter-attribute interactions

   b) $\mathbb{S} = \mathbf{I}_{K,K} - \mathbf{1}_{K,K}$
      $\leadsto$ same inter-attribute interactions and no intra-attribute interactions

   c) $\mathbb{S} = \mathbf{1}_{K,K}$
      $\leadsto$ full agreement between attributes.

3. $\mathbf{\Omega}$ is the nearest a positive definite approximation of $\mathbf{M}$
4. Control the difficulty with $\gamma > 0 : \mathbf{\Omega} = \mathbf{\Omega} + \gamma I$;
5. Draw an i.i.d. n-size sample $\mathbf{X} \in \mathbb{R}^{n \times pK}$ of $X \sim \mathcal{N}\left(0, \mathbf{\Omega}^{-1}\right)$.

Example: original graph + intra/inter/full aggrements



Simulation study: evaluation

Competitors

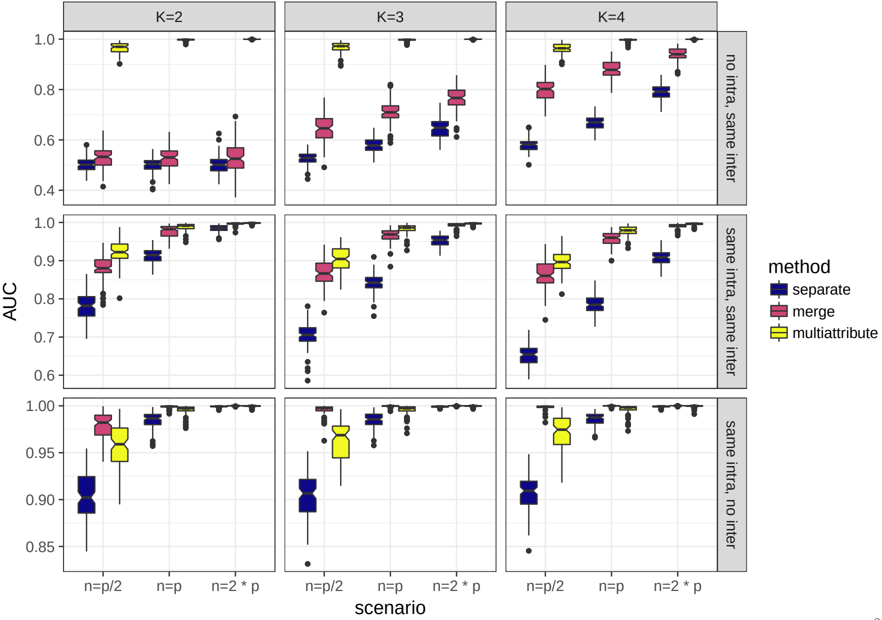
- `multiattribute`: reconstruct one network with K data sets $\mathbf{X}^{(1)}, \dots \mathbf{X}^{(K)}$ all with size $\mathbb{R}^{n \times p}$
- `separate`: reconstruct K networks with K data sets $\mathbf{X}^{(1)}, \dots \mathbf{X}^{(K)}$ all with size $\mathbb{R}^{n \times p}$
- the `merge` variant: reconstruct one network by merging $\mathbf{X}^{(1)}, \dots \mathbf{X}^{(K)}$ into a single $\tilde{\mathbf{X}}$ data set in $\mathbb{R}^{nK \times p}$

Performances

Use area under ROC curve (AUC). For the *separate* variant, the retained AUC is the AUC averaged over all attributes.

$\rightsquigarrow$ Set $p = 40$, vary n, K and replicate 100 times

Simulation study: results



Breast cancer data: application

Two cohorts with both proteomic and transcriptomic data

1. NCI-60: $n = 60$ diverse human cancer cell lines, $p = 91$
2. RATHER: $n = 100$ sample from patients with breast cancer, $p = 117$

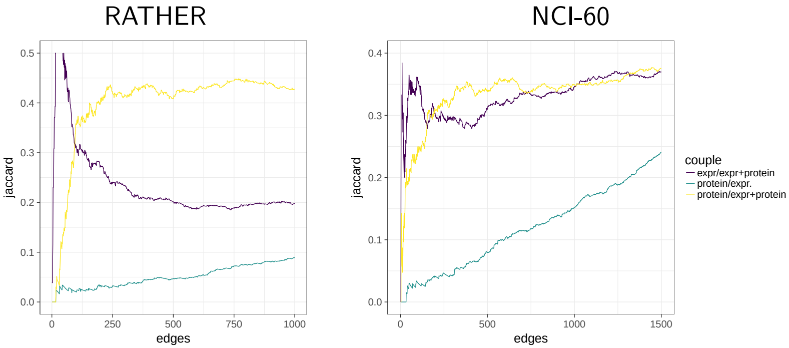


Figure: Jaccard's similarity index $J(A,B) = \frac{|A \cap B|}{|A \cup B|}$ between uni-attribute and multiattribute networks, for RATHER and NCI60 data set: multiattribute networks share a high Jaccard index with both uni-attribute networks.

Inferred networks

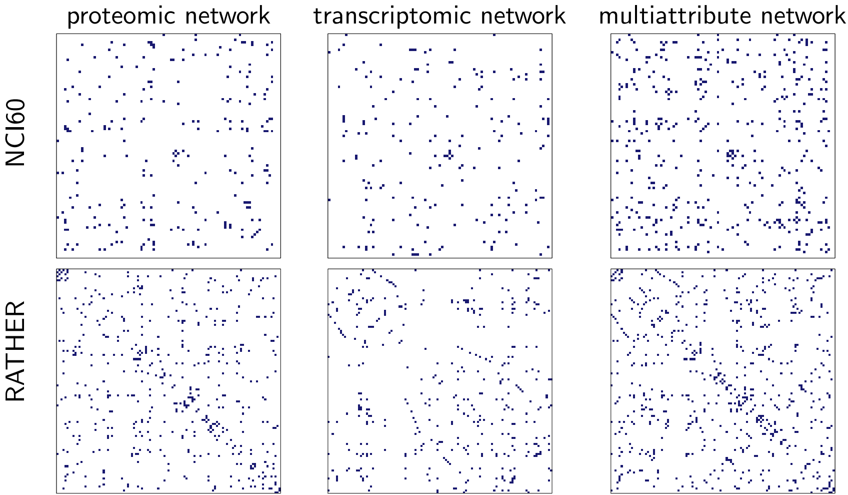


Figure: Uni-attribute and multiattribute networks inferred on both NCI60 and RATHER dataset. The number of neighbors of each entity is chosen by cross-validation. Multiattribute networks catch motif found in the uniattribute counterparts.

Conclusion

Perspectives

- Validation?
- Other penalties?
- Covariates?

Thanks to you for your patience and to my co-workers