La statistique non asymptotique au service de la génétique.

Emilie Devijver & Mélina Gallopin

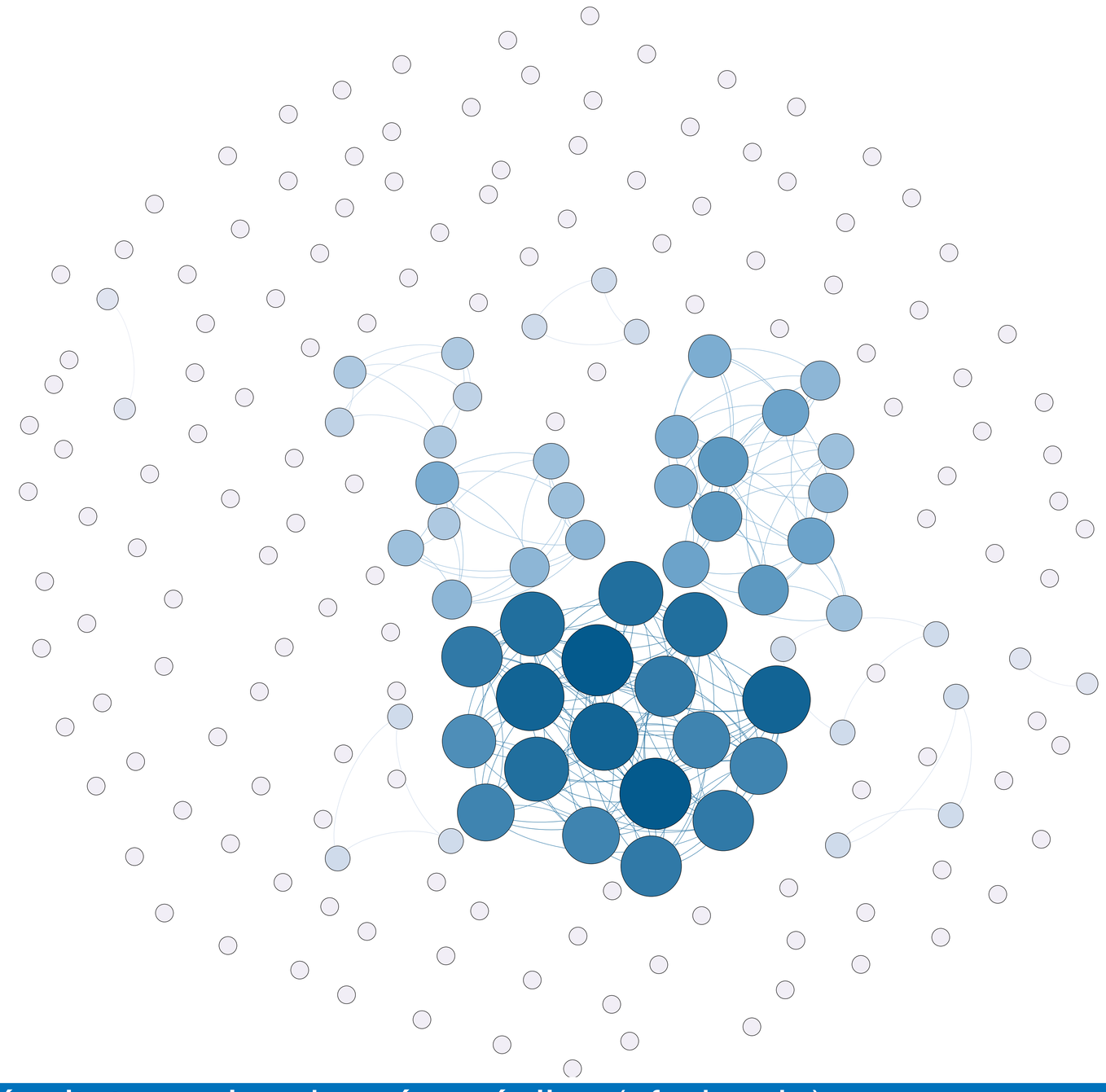
Reconstruction de réseaux d'intéractions géniques

Mieux comprendre le vivant

- Comprendre finement le fonctionnement des cellules et trouver des solutions thérapeuthiques aux maladies infectieuses et cancéreuses
- Détecter l'activité des gènes dans le génome grâce à une nouvelle technologie (RNA-seq, 2008)
- Utiliser des outils statistiques pour comprendre les mécanismes biologiques sous-jacents

L'inférence de réseaux d'intéractions entre les gènes

- Reconstituer le réseau de régulation de gènes pour détecter les gènes les plus importants dans le mécanisme biologique étudié



Résultats sur les données réelles (cf. dessin)

- 140 gènes isolés = 140 gènes n'influent pas sur les autres
- 2 blocs de taille 2, 4 blocs de taille 3
- 1 bloc de taille 18, 1 bloc de taille 13, 1 bloc de taille 8, 1 bloc de taille 5

$\Rightarrow$ taille des blocs très raisonnable pour l'analyse.
N.B. : le diamètre des sommets est proportionnel à l'importance de chaque gène.

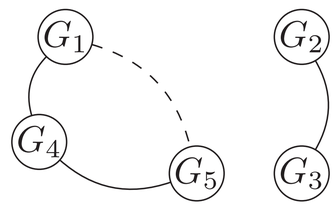
Un cadre mathématique rigoureux

Gène = sommet du graphe $\sim$ variable aléatoire

Dépendance = arête du graphe $\sim$ matrice de covariance Σ

Dépendance directe = arête pleine du graphe
$\sim$ matrice de précision Σ^{-1}

$$(G_1, \ldots, G_p) \sim \mathcal{N}_p(\mathbf{0}_p, \Sigma)$$



$$\Sigma^{-1} = \begin{array}{c} \begin{matrix} G_1 & G_2 & G_3 & G_4 & G_5 \end{matrix} \\ \begin{pmatrix} \star & 0 & 0 & \star & 0 \\ 0 & \star & \star & 0 & 0 \\ 0 & \star & \star & 0 & 0 \\ \star & 0 & 0 & \star & \star \\ 0 & 0 & 0 & \star & \star \end{pmatrix} \end{array}$$

Des données biologiques limitées

Coût financier élévé de la technologie RNA-seq
$\Rightarrow$ nombre d'échantillons séquencés limité

Jeu de données étudié (*Pickrell et al.*, 2010)

- 69 individus séquencés
- 200 gènes à étudier $\Rightarrow$19 900 paramètres à calculer

Inférence impossible !

Besoin d'une méthode non asymptotique, i.e. qui fonctionne pour une taille d'échantillon limitée

Notre approche

But

- Réduire le nombre de paramètres à estimer
- Améliorer la qualité de l'inférence
- Faciliter l'analyse pour les généticiens

Idée générale

- Construire des groupes de gènes indépendants les uns des autres
- Inférer les réseaux au sein de chaque groupe de gènes (inférence améliorée car le nombre de paramètres à estimer est réduit)

Conclusion

✓graphe facilement interprétable
✓méthode adaptée aux données réelles
✓méthode justifiée par des résultats théoriques valables avec peu d'observations (69 individus)

Références

[1] J. Pickrell. *Understanding mechanisms underlying human gene expression variation with RNA sequencing*. Nature, 7289(464) 768–772, 2010.

[2] E. Devijver & M. Gallopin. *Block-diagonal covariance selection for high-dimensional Gaussian graphical models*. Journal of the American Statistical Association, 2016, à paraître.

[3] Package R *shock*, CRAN

[4] M. Bastian, S. Heymann & M. Jacomy *Gephi: An Open Source Software for Exploring and Manipulating Networks*. International AAAI Conference on Weblogs and Social Media, 2009.

Garanties théoriques

Méthodologie (procédure shock)

- Construction de plusieurs graphes définis par Σ^{-1} plus ou moins denses
- Sélection d'un graphe par un critère approprié à notre cadre, avec des garanties théoriques

Justification du critère de sélection du graphe

Théorème (inégalité oracle)
Le graphe choisi est aussi bon que l'*oracle*, le meilleur graphe parmi ceux que l'on a construits[a].
Si on note $f^\star$ la densité associée au vrai graphe, $\hat{\Sigma}$ le graphe qu'on sélectionne, C une constante universelle, et d une distance appropriée,

$$d(\mathcal{N}(0, \hat{\Sigma}), f^\star) \leq C \min_{\Sigma \text{ construits}} d(\mathcal{N}(0, \Sigma), f^\star).$$

Théorème (borne minimax)
En toute généralité, sans plus d'hypothèses sur la structure, le critère de sélection est optimal.

[a] On ne connaît pas l'oracle, car il dépend du vrai graphe, que l'on ne connaît pas en pratique !