RESYS, PROJET 7: Réseau épigénétique

Alex YE

Objectif

- Construire un réseau de régulation
  - Réseau de régulation
    - Mark-Target
    - Causal/Orienté
  - Méthodes d'inférence
    - HC: Score-based
    - PC: Constraint-based
    - miic: Constraint-based, basé sur PC, pas de contrôle sur la densité du réseau.
    - MMHC: Hybrid, (constraint-based skeleton + score-based orientation)
    - GENIE3: Feature importance
  - Quantifier les résultats
    - Fold enrichment
    - Heuristiques

Données

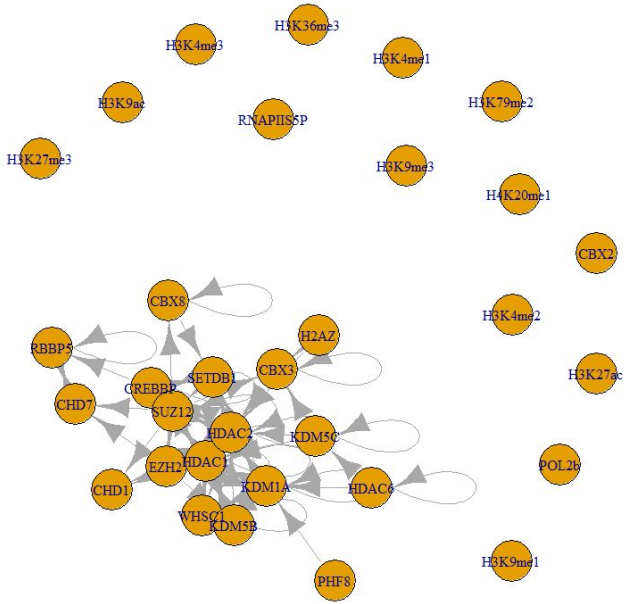
TF: facteur de transcription, gènes présents dans les données d'expression

- Données d'expression génétique (INPUT, given)
- Données expérimentales (SCORE, online)
  - BIOGRID
    - 508019 interactions (240 entrées liées à nos TF)
    - Pas de donnée sur les interactions entre histones
    - https://thebiogrid.org/
  - HI-union
    - 64006 interactions (0 entrées liées à nos TF)
    - http://interactome.baderlab.org/
  - HuRI
    - 52570 interactions (0 entrées liées à nos TF)
    - http://interactome.baderlab.org/

Réseau référence (données expérimentales)

32 noeuds
68 arcs
12 autorégulations
13 noeuds inactifs

#arc moyens = 3.58
#arcs estimé = 114.53



Fold enrichment

Fonction de score qui cumule erreur de type I et type II

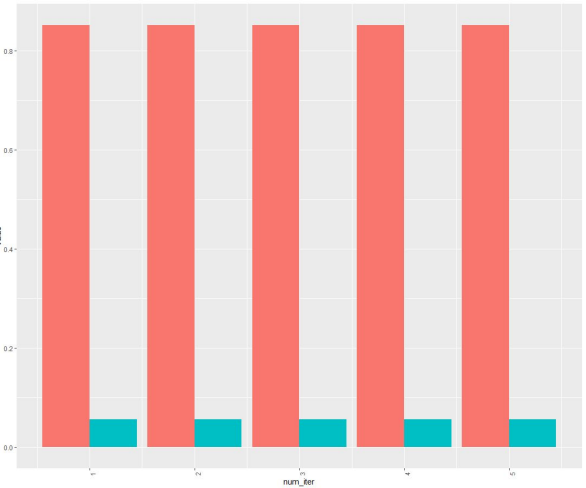
$$Score_{fe} = \frac{\#correct\_edges}{\#randomly\_correct\_edges}$$

$$Score_{fe} = \frac{TP}{\frac{(\#network\_edges).(\#dataset\_edges)}{N}}$$

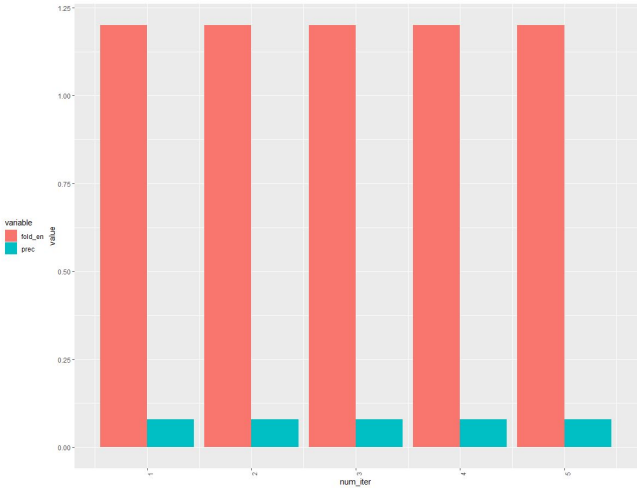
$$Score_{fe} = \frac{TP.N}{(TP+FP).(TP+FN)}$$

TP: arc prédit et présent dans le réseau de référence.
FP: arc prédit mais absent du réseau de référence.
FN: arc absent du réseau de référence.
N: nombre d'arc total (32x32)

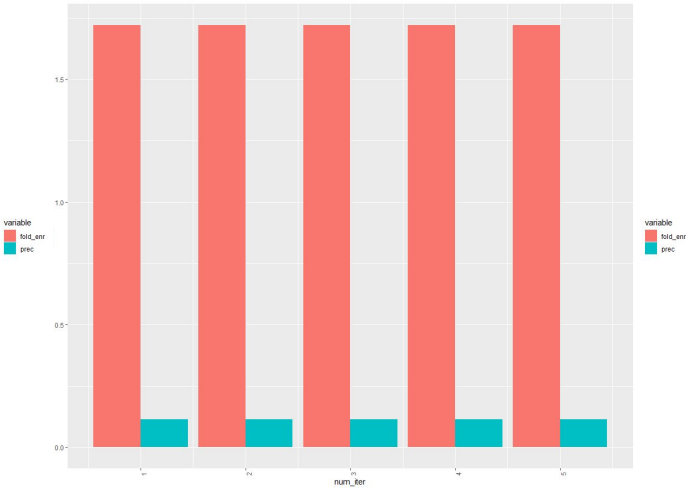
Fold enrichment et Précision sur 5 appels



HC



GENIE3



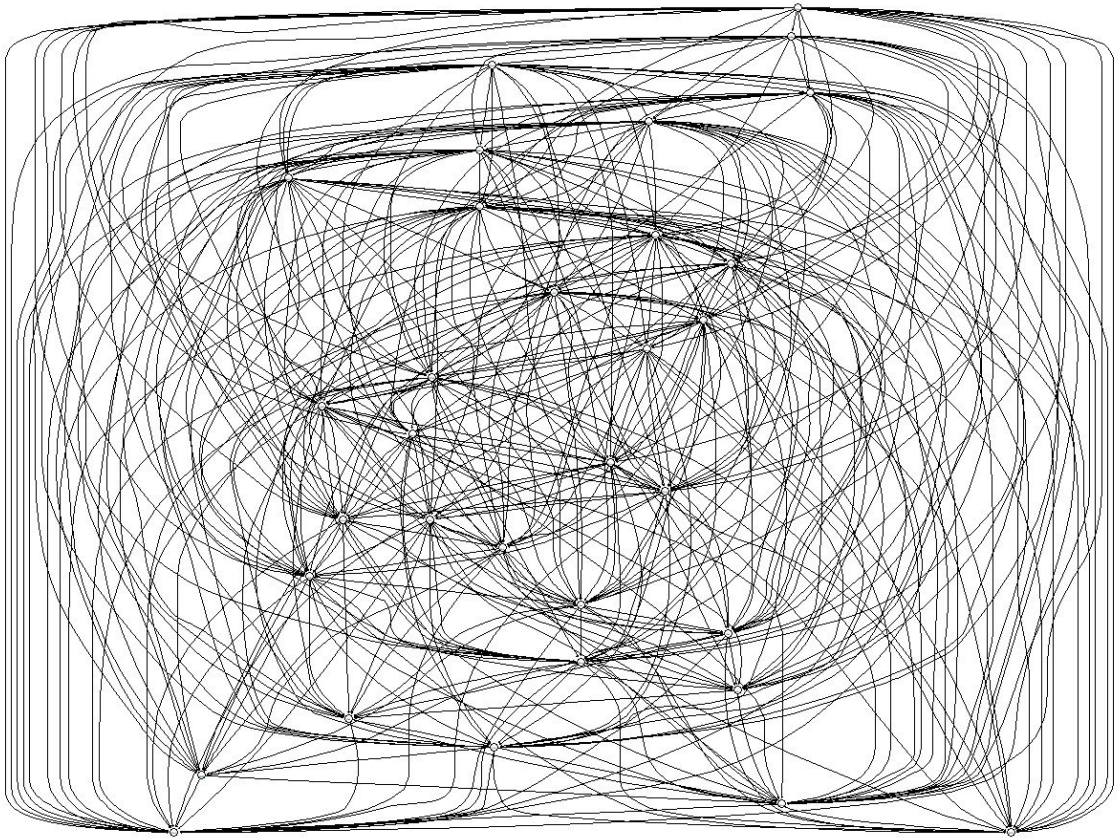
miic

Remarque: Aucun changement entre 2 appels, il n'y a pas d'aléa dans l'initialisation.

Fold enrichment, HC

0 autorégulateur
389 arcs
fold enrichment: 0.8517

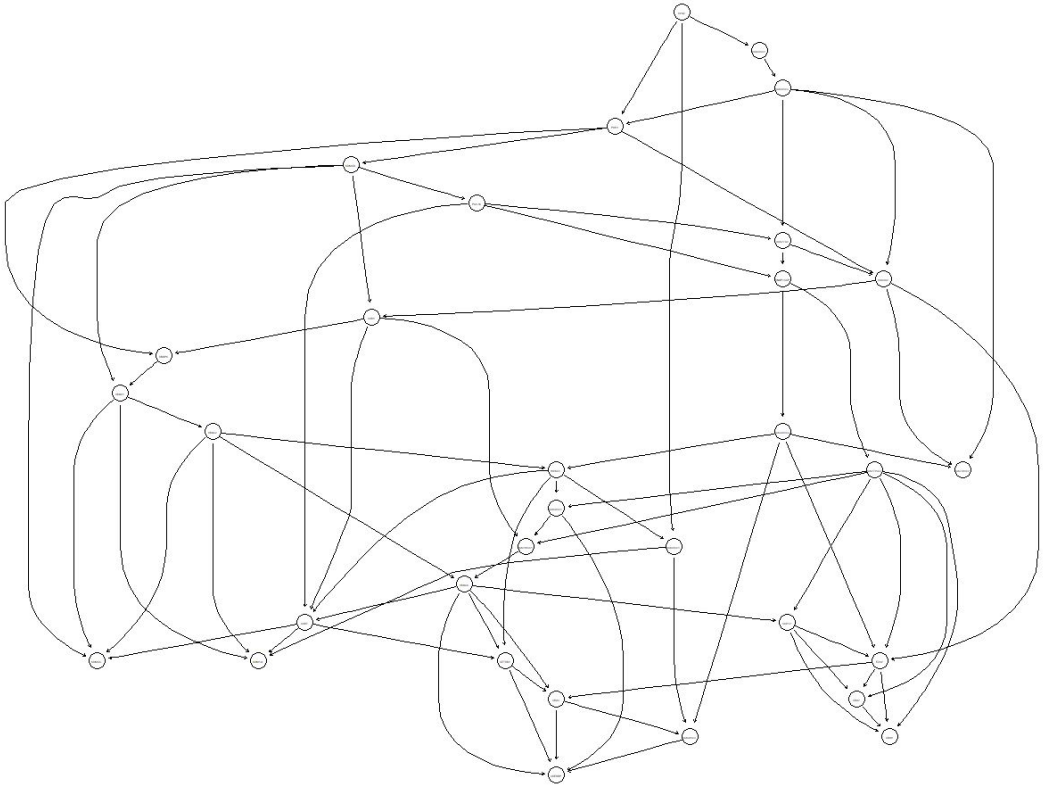
#arcs estimé = 114.53



Fold enrichment, MMHC (alpha: 0.05)

0 autorégulateur
75 arcs
fold enrichment: 1.3514

#arcs estimé = 114.53

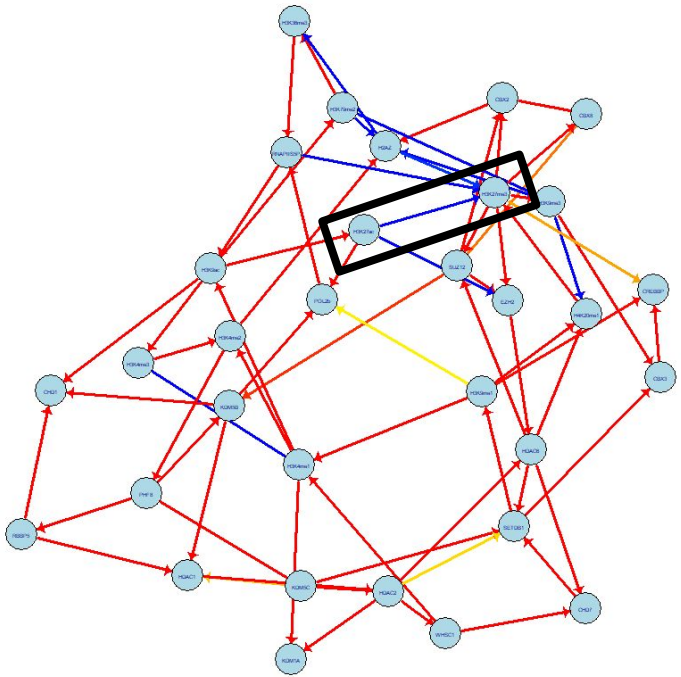


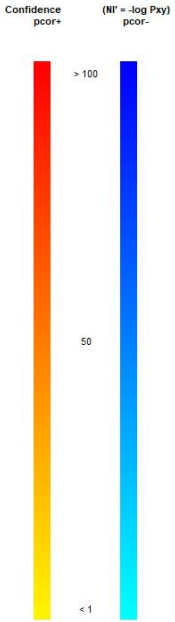
Fold enrichment, miic

arc(H3K27ac → H3K27me3) à faible indice de confiance
2 TF incompatibles, l'algorithme ne fait pas n'importe quoi

0 autorégulateur
140 arcs
fold enrichment: 1.721

#arcs estimé = 114.53

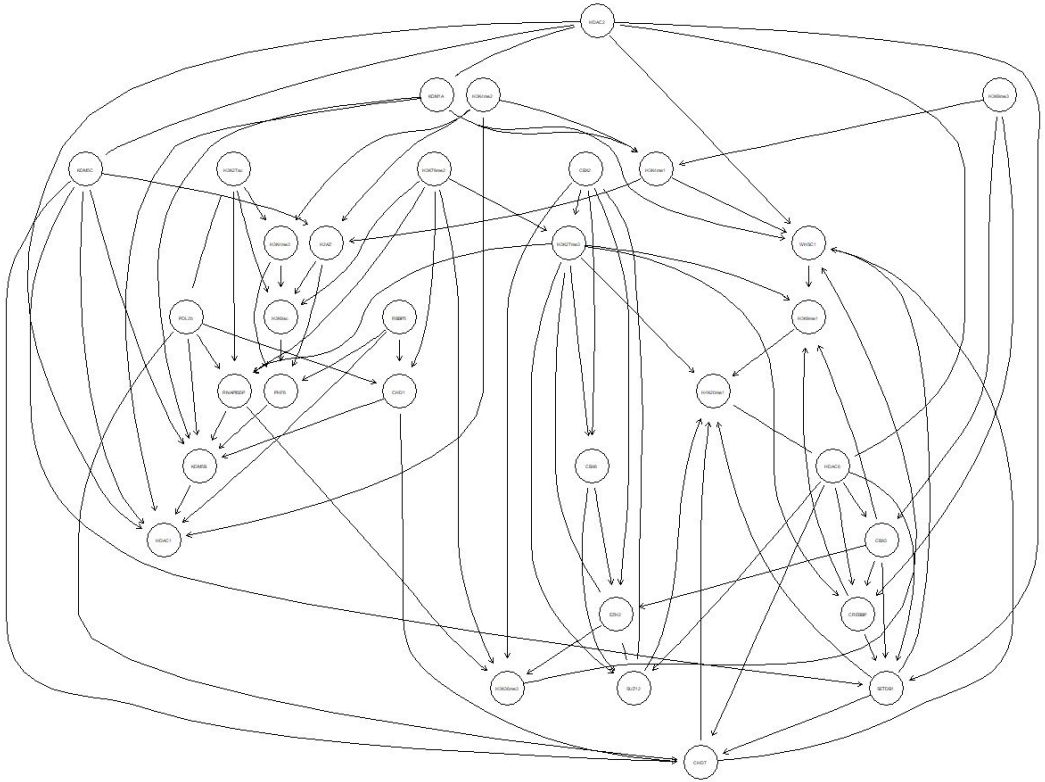




Fold enrichment, PC (alpha: 0.4)

0 autorégulateur
96 arcs
fold enrichment: 2.039

#arcs estimé = 114.53

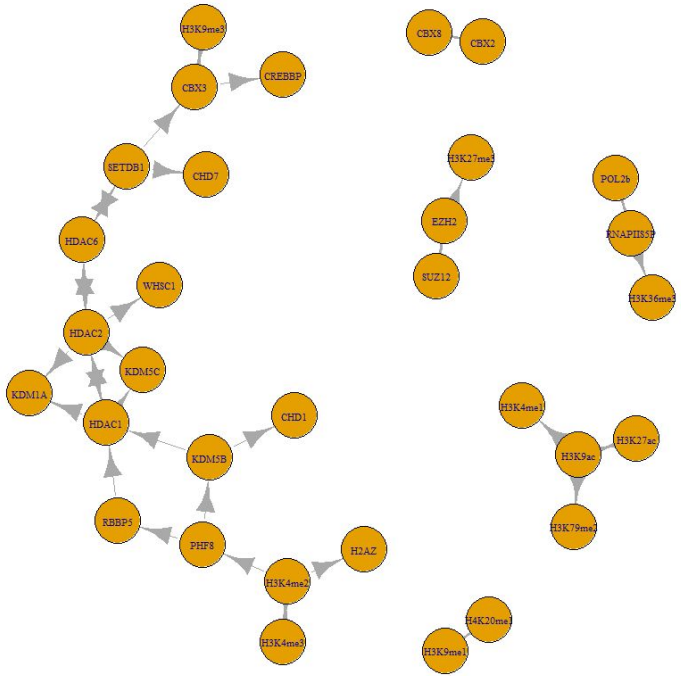


Fold enrichment, GENIE3 Random Forest

6 composantes

0 autorégulateur
39 arcs
fold enrichment: 3.088989

#arcs estimé = 114.53

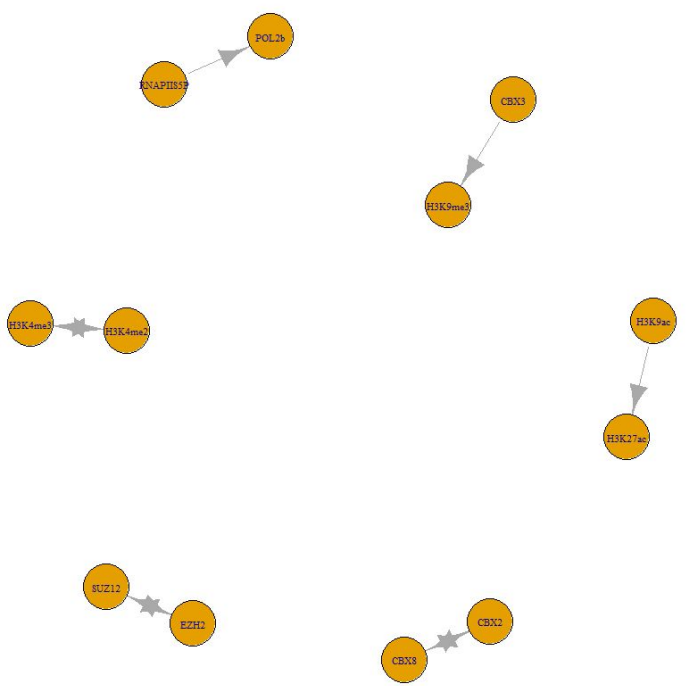


Fold enrichment, GENIE3 Extra Trees

Il n'y a presque rien

0 autorégulateur
9 arcs
fold enrichment: 3.346405

#arcs estimé = 114.53



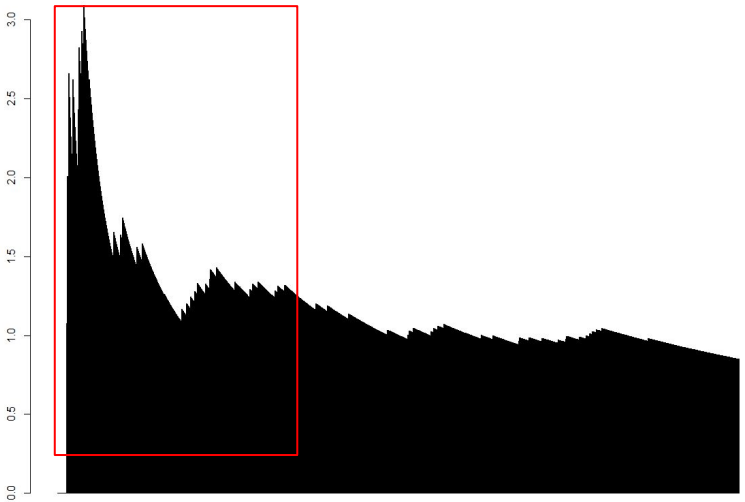
Remarques

- Aucun arc d’autorégulation.

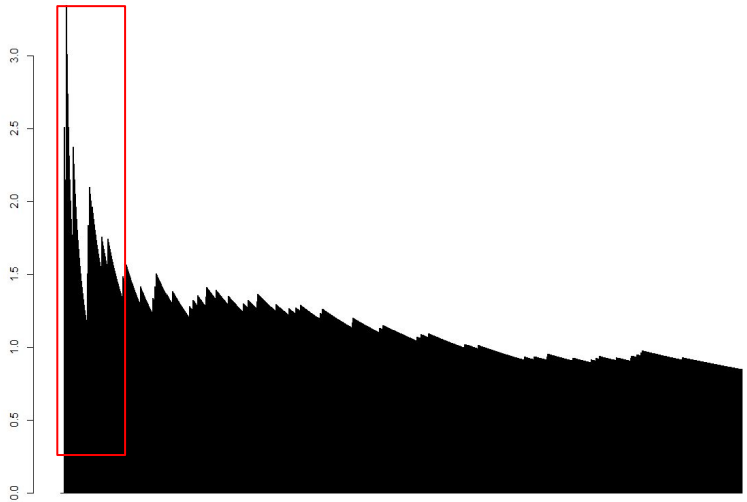
- Résultats de GENIE3 révèlent peu d’information tout en ayant les meilleurs scores. L’un a la moitié des interactions observées expérimentalement (39/68) et l’autre en a seulement (9/68).

  Intuition: Peu probable que ce soient réellement les optimums, le fold enrichment est inadapté pour l’algorithme GENIE3.

Fold enrichment, GENIE3 en fonction du #arc



RF



ET

Remarque: Dû au coût croissant du nombre d'arc, une région qui ne trouve pas de TP(vrai positif) créer un "fossé" et diminue les gains pour les prochains TP, ce qui fait que la seconde "colline" sera plus basse que la première. Ce qui est étrange c'est que le fossé existe chez les arcs de (très) haute confiance.

Pénalité

Nous partons du principe que l'algorithme GENIE3 ne fait pas n'importe quoi, et que le "fossé" est causé par un manque d'information expérimentale (utilisé pour détecter un TP).
Nous chercherons donc à accéder à la seconde "colline".
Pour cela nous estimons le nombre d'arc du réseau en utilisant les données expérimentales présentes puis nous calculons l'écart entre le nombre d'arc estimé et le nombre d'arc du réseau prédit.

Ensuite nous l'incoporons dans la fonction de scoring pour optimiser nos réseaux.
Plus l'écart est grand plus la pénalité est grande.

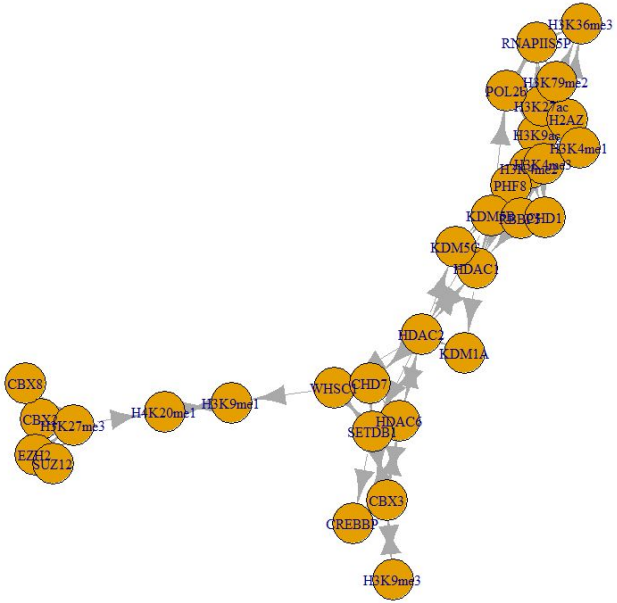
$$Score_{penalty} = \frac{n_{arc\_est}}{1 + (n_{arc\_pred} - n_{arc\_est})^2} . Score_{fe}$$

Remarque: Lors du choix entre plusieurs réseaux, le score de fold enrichment garde la priorité.

Pénalité, GENIE3 Extra Trees

0 autorégulateur
115 arcs
fold enrichment: 1.309
penalty score: 122.83

#arcs estimé = 114.53

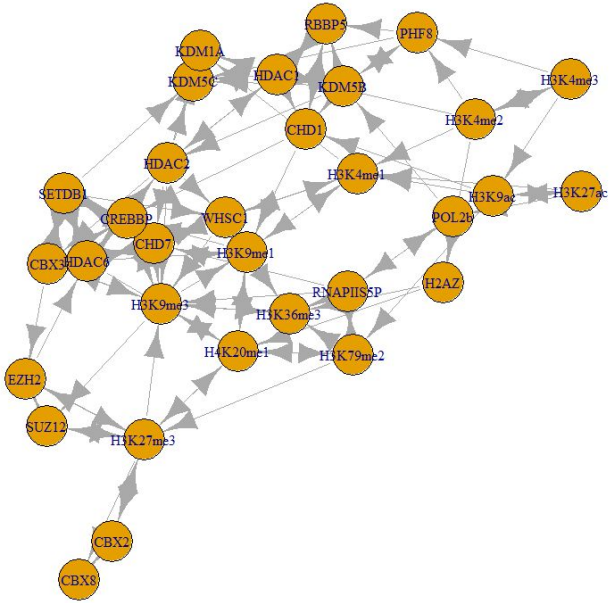


Pénalité, GENIE3 Random Forest

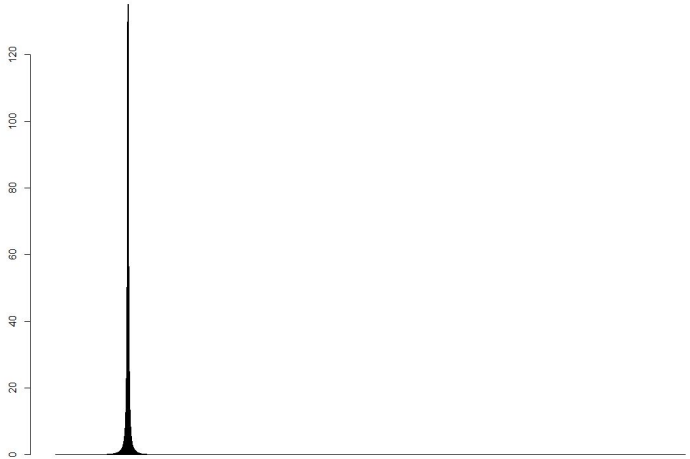
Meilleur en tout point à Extra Trees selon nos métriques.

0 autorégulateur
115 arcs
fold enrichment: 1.440
penalty score: 135.12

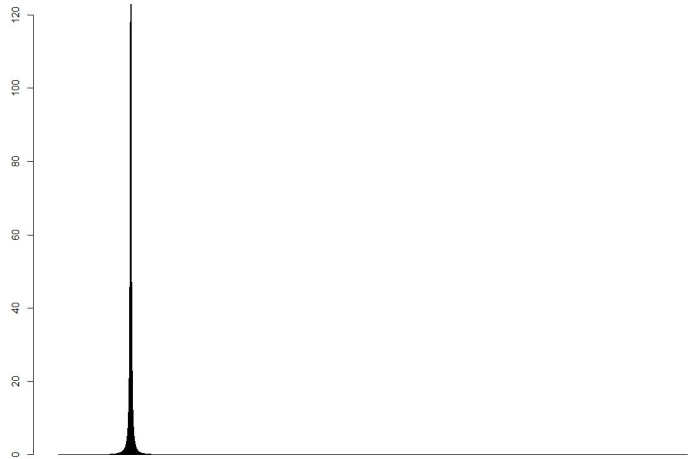
#arcs estimé = 114.53



Score de pénalité en fonction du #arc



RF



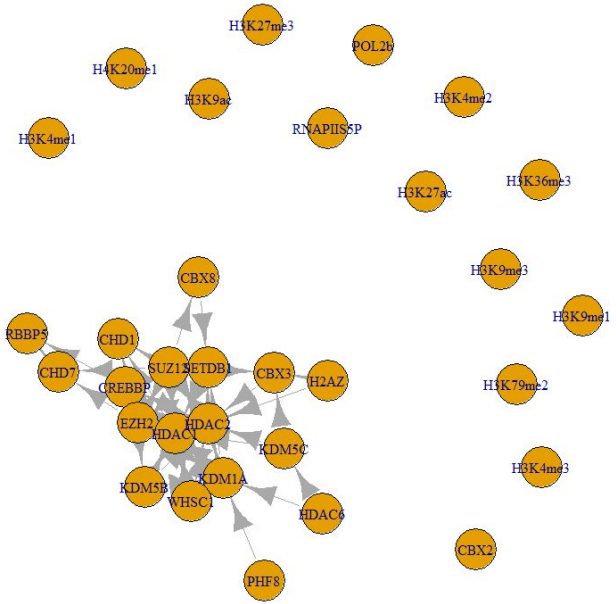
ET

Remarque: Biais énorme sur le nombre d'arc estimé, l'optimum est exactement au nombre d'arc estimé (possible mais peu probable). Un peu l'inverse du premier problème.

Réseau référence, sans autorégulation

32 noeuds
56 arcs
0 autorégulations
13 noeuds inactifs

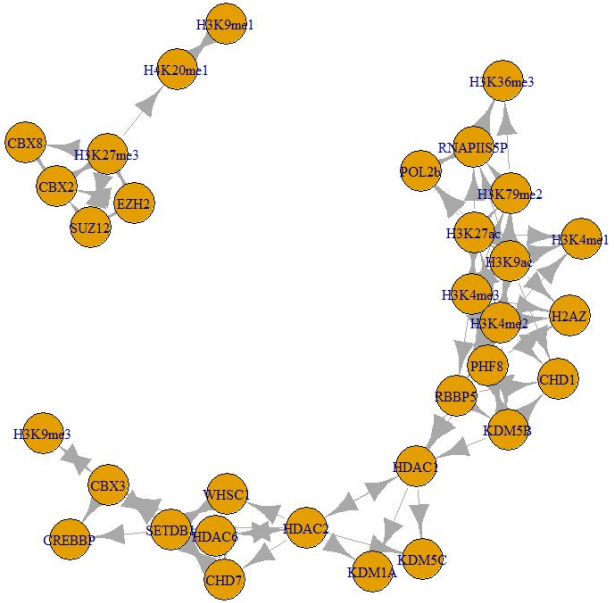
#arc moyens = 2.947
#arcs estimé = 94.315



GENIE3 Extra Trees

0 autorégulateur
97 arcs
fold enrichment: 1.441
penalty score: 123.653

#arcs estimé = 94.315

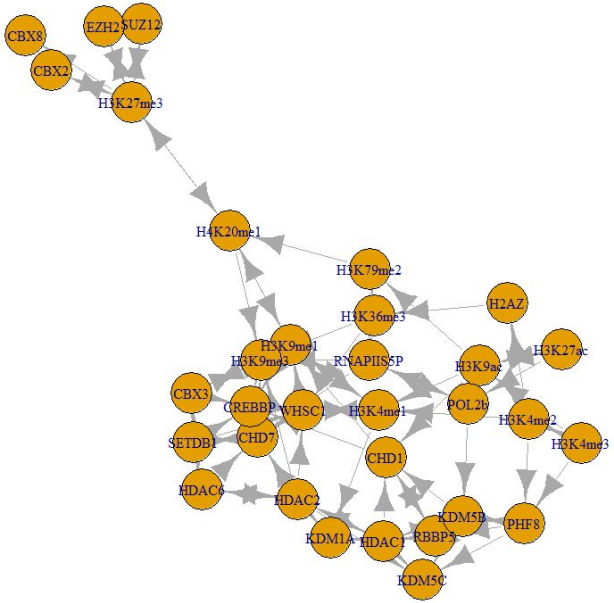


GENIE3 Random Forest

Meilleur en tout point à Extra Trees selon nos métriques.

0 autorégulateur
97 arcs
fold enrichment: 1.602
penalty score: 137.392

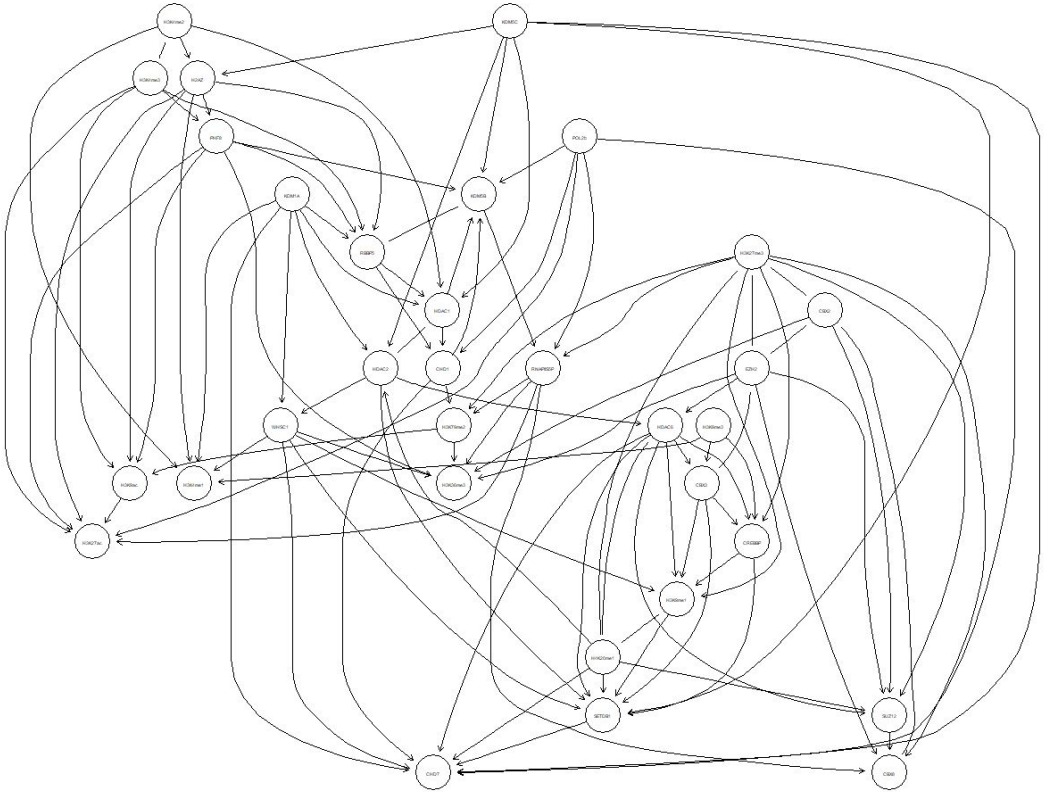
#arcs estimé = 94.315



PC (alpha: 0.1)

0 autorégulateur
97 arcs
fold enrichment: 1.902
penalty score: 122.2

#arcs estimé = 94.315



Remarques

Pour PC, le résultat n’est pas si différent que lors de l’optimisation seulement sur Fold enrichment, point positif car il n’était pas sensé être affecté.

TOP 3:

1. PC (alpha=0.1), 1.902
2. miic, 1.721
3. GENIE3 (Random Forest), 1.602

Conclusion

Les méthodes d’inférence constraint-based et feature importance sont les plus adaptées à cette tâche.

Nous réussissons à générer des réseaux vraisemblables en utilisant des heuristiques pour réduire la dépendance aux données expérimentales.

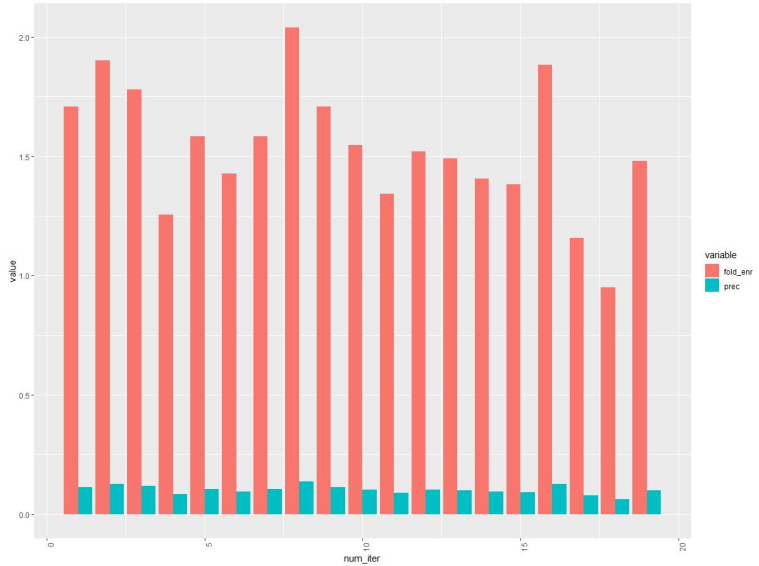
Une augmentation de la quantité de données expérimentales disponibles nous permettra d’obtenir des résultats plus précis. Surtout pour les données histones, car nous n’en avons aucune.

Questions en suspens

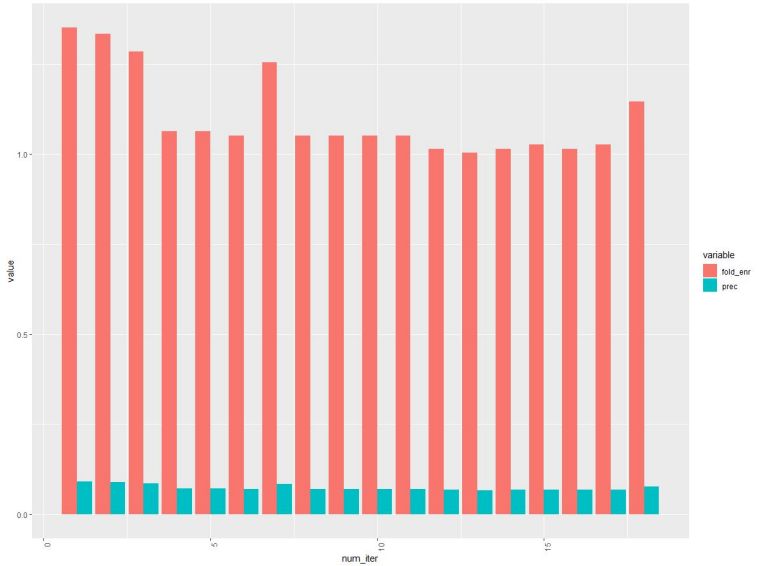
- Quelle fiabilité peut-on accorder à ces résultats?
  - PC: Les résultats en fonction de alpha sont désordonnés
  - Le nombre d’arc estimé semble résoudre notre problème, et sur d’autres problèmes?

Quelles probabilité que nos résultats obtenus soient dus simplement à la chance?

Fold enrichment, PC et MMHC en fonction de alpha



PC



MMHC

Remarque: On ne peut pas en dire grand chose, on s'attendait plutôt à un maximum à gauche et un minimum à droite

Pénalité, PC (alpha: 0.8)

alpha est très élevé, peut-on se fier à ces résultats ?

0 autorégulateur
112 arcs
fold enrichment: 1.882
penalty score: 29.13

#arcs estimé = 114.53