Université Paris Descartes

MLSD 2 2018-2019

Méthodologie de la recherche

Professeur :
Séverine Affeldt
LIPADE

Auteurs :
Joseph Gesnouin
Yannis Tannier

Contents

1 Introduction

Le but de ce projet est de reconstruire un réseau de gênes et de l'analyser afin de mieux comprendre l'impact de certains facteurs de transcriptions dans l'expression des gènes et donc le développement cellulaire de certaines cellules souches spécialisées.

Les facteurs de transcriptions sont des protéines nécessaires à l'initiation ou à la régulation de la transcription d'un gène dans l'ensemble du vivant. la transcription est un mécanisme qui permet de recopier les données des gènes, ce qui permet leur utilisation pour créer de la matière biologique en assemblant des acides aminés en protéines selon leur code génétique.

Suite à l'expression de certains facteurs de transcriptions lors de son developement, une cellule sera plus à même de remplir certains rôles au niveau du vivant.

Ce rapport s'inscrit dans cette volonté de regrouper et de classer les liasons inter-facteurs afin de déterminer quels facteurs participaient au développement de chacun des types de cellules proposés.

Celui-ci se divise en trois parties: dans un premier temps, nous avons cherché à parfaire nos connaissances sur le jeu de données fourni mais également regrouper les cellules disponibles dans le jeu de données en sous-groupes en fonction des facteurs de transcription s'étant exprimés durant leur développement grâce à des procédés d'apprentissage automatique: NMF & Spherical K-means. Nous avons ensuite généré des visualisations de ces liaisons via des algorithmes de réduction de dimensions comme t-SNE.

Dans un second temps, nous avons souhaité recréer le réseau de gênes et analyser les résultats obtenus, puis pour terminer, nous avons souhaité comparer nos résultats à l'état de l'art disponible comme l'algorithme miic de construction de réseaux: Multivariate Information based Inductive Causation

2 Exploration des données

Le jeu de données sur lequel nous avons travaillé contient l'expression de 33 facteurs de transcriptions pour 3934 cellules dont les données ont été extraites lors de quatres periodes distinctes de leur développement. Notre matrice correspondait à une concaténation de l'expression de ces facteurs de transcriptions pour les 3934 cellules sans prendre en compte la periode d'expression du gène mais seulement si celui-ci c'était exprimé à un certain moment du développement de la cellule. Nous avons souhaité avant toute tentative de modélisation de graphe, d'en apprendre plus sur le jeu de données qui nous avait été fourni.

2.1 Spherical K-means

La matrice pouvant s'apparenter à une matrice document-terme binaire étant sparse à 45%, nous avons souhaité approximer le nombre de classes des cellules afin d'identifier des comportements récurrents dans l'expression des facteurs de transcription chez les cellules du jeu de données.

[**1**] avait utilisé une classification de type K-means afin de différencier ces cellules en trois classes distinctes. Nous avons fait le choix d'utiliser une variante de K-means pour déterminer le nombre de classes: Sphérical K-means.

Grâce au modèle vectoriel, il est possible d'arriver à quantifier la proximité sémantique entre chaque cellule en introduisant des mesures de comparaisons plausibles, principalement via la distance angulaire: : deux cellules ayant globalement les mêmes gênes s'étant exprimés lors de son développement auront généralement la même direction dans l'espace vectoriel.

Contrairement à l'algorithme K-means, la fonction de minimisation de Spherical K-means ne prend pas en compte la distance euclidienne mais la distance angulaire, grâce au cosinus:

$$\sum_{i}(1 - \cos\left(x_{ij}p_{c(i)}\right))$$

En choisissant de représenter notre matrice comme un matrice document-terme(ie maladie-facteur de transcription) sous le format vectoriel, il faut inévitablement choisir une distance angulaire afin de les comparer. Il s'avère que cette distance propose de bien meilleurs résultats que les autres distances puisqu'elle conserve généralement mieux la valeur sémantique des textes et n'est pas affectée par leur

longueurs. En effet, pour classifier des documents textuels, il est plus raisonnable de comparer les facteurs de transcription qui s'expriment plutôt que le nombre de facteurs s'étant exprimés lors du développement de la cellule.

Après un tf-idf et afin de déterminer le nombre de clusters à garder sur toutes nos valeurs nous nous sommes inspiré de la methode d'Elbow: il s'agit de tracer la courbe entre le nombre de clusters et la fonction de coût afin de déterminer un "coude". Nous avons donc réalisé des spherical K-means pour des rangs allant de 2 à 15 afin de savoir à quelle valeur de k nous allions nous intéresser.

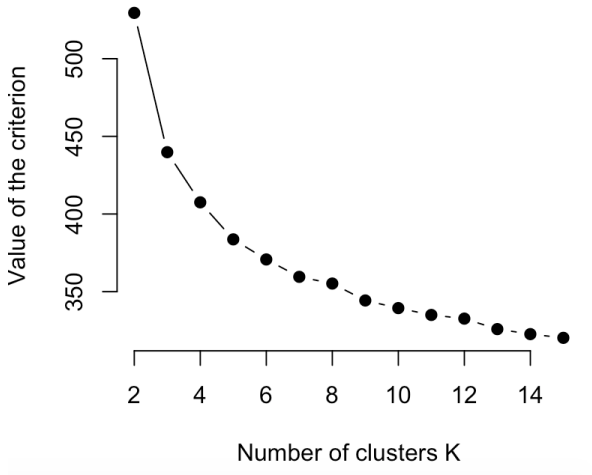

Figure 1: Résultat de la méthode d'Elbow sur spherical k-means : nombre de cluster variant de 2 à 15

Cette méthode regarde le pourcentage de variance qu'apporte chaque cluster. Nous déterminons le nombre de clusters à choisir au moment où rajouter un autre cluster n'apporte pas réellement d'information pour modéliser les données. En règle générale, les premiers clusters apporteront énormément d'information car ils expliquent beaucoup de variance, mais à un certain moment le gain marginal devient négligeable, ce qui donne un angle au graphe. Le nombre de clusters est déterminé par ce point. Dans notre cas, le coude n'était pas flagrant, nous avons donc eu recours à un calcul pour trouver le rang k : nous avons considéré que ce rang était défini comme le point avec la plus grande distance orthogonale depuis les extrémités. Nous trouvons un nombre de clusters de 3 avec la méthode Spherical K-means. Ce résultat est en adéquation avec le nombre de clusters trouvés par [**1**].

Notre clustering nous aura permis de départager ces 3934 cellules en trois groupes distincts en fonction des gènes s'étant exprimés lors de leur développement. Nos trois clusters sont plus ou moins équilibrés en taille avec respectivement 1021, 735 et 2178 cellules dans chacun: on remarque cependant un sous-groupe représentant environ la moitié des cellules du jeu de données.

2.2 NMF

Nous avons souhaité réaliser une factorisation matricielle non négative en initialisant W_0 grâce aux centroïdes générés par le clustering préalablement réalisé via l'algorithme spherical k-means.

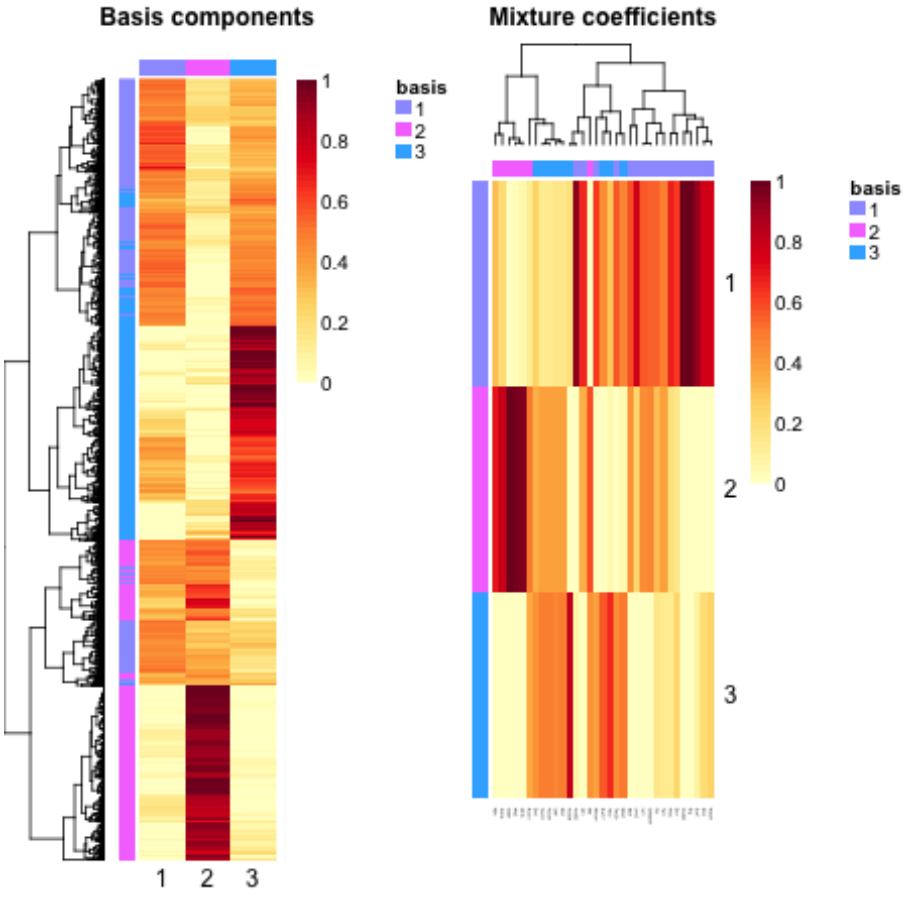

Figure 2: Heatmap de la matrice basis et de la matrice coefficient

Nous remarquons bien cette differentiation du type des cellules en fonction des ex-

pressions des transcripts mais également un regroupement de certains gênes s'étant exprimés lors du développement des cellules, qui seront alors utiles à la caractérisation de celles-ci.

2.3 t-SNE

Dernière étape d'analyse initiale de notre jeu de données, nous avons souhaité visualiser et représenter graphiquement dans un espace de dimension réduit nos cellules grâce à l'algorithme de réduction de dimension t-SNE en y ajoutant les informations des groupes trouvés précédement:

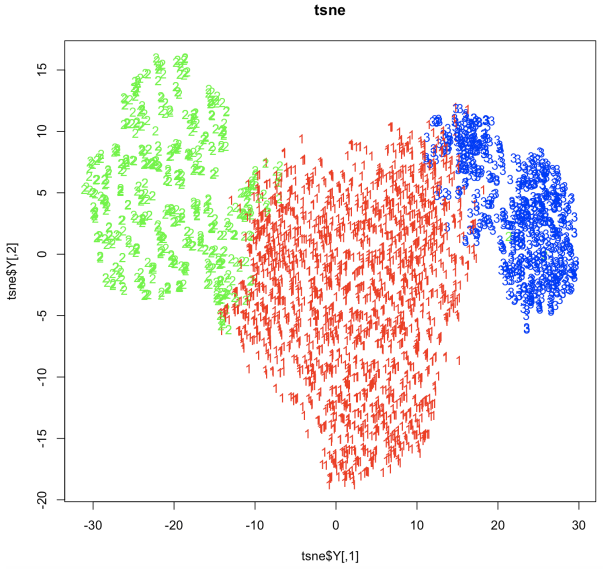

Figure 3: t-SNE

Grâce à tout ces pré-traitements et notre clustering en adequation avec ce que nous avions lu dans [**1**], nous pouvons en conclure que des 3934 cellules proposées, celles-ci peuvent se séparer en trois sous-groupes grâce à une classification de leur fluctuation d'expression de gênes: progénitrices, précurseuses endothéliales, et précurseuses hématopoïétiques.

Afin de parfaire notre analyse nous avons cherché à comprendre un peu mieux ce que à quoi ces sous-groupes correspondaient:

- **endothéliales** : ces cellules tapissent la face interne des vaisseaux. Ce sont des cellules qui sont en contact direct avec le sang et qui assure l'intégrité des

vaisseaux. Elles sont capables de s'adapter à de fortes pressions, notamment en ce qui concerne les cellules proches de la région cardiaque. Ce sous-groupe rassemble donc les TF pour lesquels nous avons trouvé une relation dans la formation de cellules endothéliales. Les genes que nous avons trouvés caractérisant ce sous groupe sont: Ets1, Erg, Etv2, Tbx3, Tbx20 et HoxB4.

- **hématopoïétique** :ces cellules sont à l'origine de toutes les lignées de cellules sanguines, celles-ci s'engagent dans une voie de différenciation qui pourra, à terme, donner une cellule sanguine mature. Ces cellules se retrouvent uniquement dans la moelle osseuse. Ce sous-groupe rassemble donc les TF pour lesquels nous avons trouvé une relation dans la formation de cellules hématopoïétiques sans trouver une fonction évidente de role dans la formation des celulles endothéliales. Les genes que nous avons trouvés liés à ce sous-groupe d'après notre clustering sont: Nfe2, Myb, Mitf, Ikaros, Gfi1b, Gfi1 et Gata1.
- **progénitrices** : Ce sous-groupe de TF correspond aux gènes ayant participé aux deux différentiations précédentes, celui-ci fait donc figure de groupe "poubelle", ce sous-groupe est très visible sur notre résultat de NMF représentant le groupe majoritaire. Les genes que nous avons trouvé liés à ce sous-groupe sont: Fli1, Tal1, Cbfa2t3h, Lyl1, Sfpil1 et Hhex.

3 Création des graphes

Notre travail d'exploration des données réalisé et notre compréhension du fonctionnement des cellules souches au niveau du vivant étant perfectionnée, nous avons souhaité reconstruire le réseau de gênes associé, selon plusieurs méthodes et critères afin de correspondre au mieux au graphe présenté dans [**1**].

3.1 TF Correlation network

La première méthode que nous avons utilisé était de calculer l'ensemble des corrélations deux à deux selon le critère de Pearson des variables centrées réduites du jeu de données:

$$corr(X, Y) = r_{XY} = \frac{s_{XY}}{s_X s_Y}$$

Figure 4: Formule du critère de Pearson

Nous obtenons les corrélations suivantes pour les 33 facteurs de transcriptions:

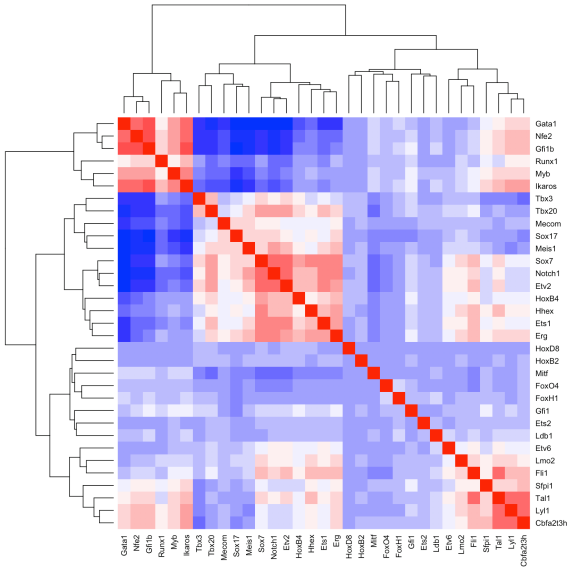

Figure 5: Formule du critère de Pearson

À première vue, il est facile de remarquer que la majorité de ces corrélations sont faiblement négatives ou faiblement positives. Afin de conserver un graphique lisible et porteur d'information nous avons selectionné une valeur seuil de corrélation afin de ne pas représenter la totalité des corrélations faibles sur le graphe:

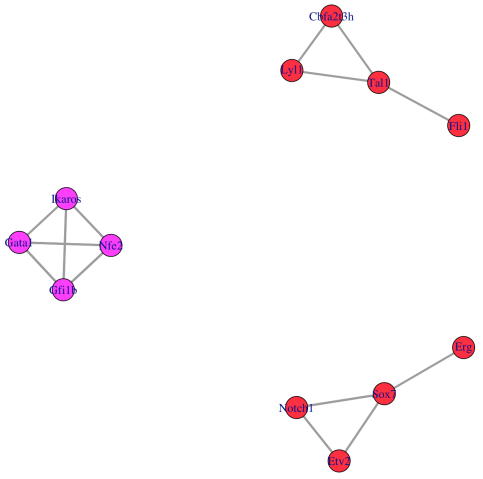

Figure 6: TF correlation graph, seuil=0.6

Le seuil fixé étant bien trop élevé pour obtenir un graphe connexe, nous remarquons ici que l'on dispose d'un graphe avec trois composantes connexes distinctes représentant les gênes s'exprimant le plus pour les trois clusters. La valeur seuil étant bien trop élevée, nous avons ensuite souhaité la réduire pour obtenir un graphe connexe porteur d'information.

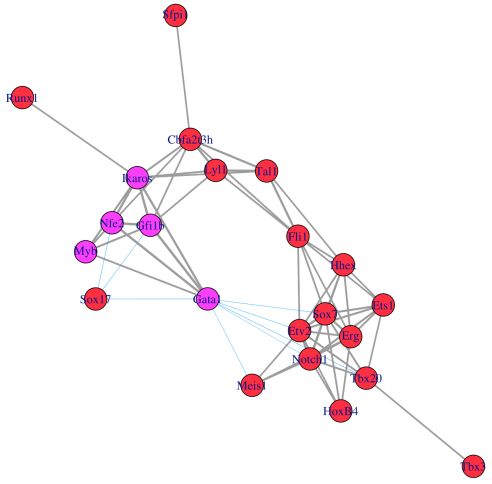

Figure 7: TF correlation graph, seuil=0.4

Ce graphique nous permet de clairement voir les facteurs étant discriminants vis à vis de l'expression d'autres facteurs, et donc à terme, du developpement de la cellule. Ainsi, nous pouvons noter l'impact notable qu'auront l'expression des TF Sox17, Gata1 dans le développement et l'utilité future de la cellule.

En conservant la même matrice des corrélations précédement calculée, nous avons souhaiter augmenter le seuil afin de voir apparaitre sur notre graphique les trois sous-groupes déterminés via notre NMF sphérical K-means:

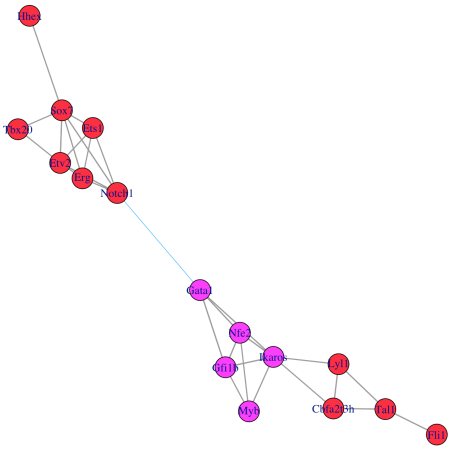

Figure 8: TF correlation graph, seuil=0.5

De ces deux graphes nous pouvons remarquer que nos résultats sont en adéquation avec le réseau du papier initial: nous trouvons bien trois groupes distincts et en comparant nos résultats, nous remarquons que nos repressions, bien qu'en nombre réduit comparé à leur graphe portent bien sur les mêmes TF: ainsi de cette première méthode nous voyons clairement que les repressions de certains gênes sont directement liées à Sox17 et Gata1.

D'un point de vue biologique, ce graphe pourrait être grandement amélioré: premièrement celui-ci est non dirigé, et l'expression d'un TF pourrait freiner le taux d'expression d'un autre TF sans que l'inverse soit vrai, ceci n'est pas pris en compte lors de cette analyse.

Dans un second temps, ce graphe pâtit du fait de ne pas prendre en compte les coefficients de corrélation partielle du jeu de données. Il s'agit d'une corrélation restant entre deux variables après avoir contrôlé (par exemple, partiellement exclue) une ou plusieurs autres variables. En effet, si l'influence de certains TF est supprimée de l'équation, certaines des corrélations précédement calculées pourraient disparaitre voir être totallement supprimées.

3.2 TF Partial Correlation network

C'est dans ce but qu'est née l'idée de créer un second réseau prenant cette fois en compte les corrélations partielles. Les couleurs des gènes sur le graphe étants relatives au clustering de la première partie.

Pour ce faire, nous avons calculé et inversé la matrice de covariance du jeu de données. L'inversion d'une matrice étant impossible si le déterminant d'une matrice est nul, nous avons du rajouter un bruit sur la diagonale de notre matrice de covariance. Afin de voir si notre bruit avait une grande influence sur notre réseau, nous avons réitéré l'opération pour plusieurs tailles de bruit différentes:

$$\mathrm{Var}(\vec{X}) = \begin{pmatrix} \mathrm{Var}(X_1) & \mathrm{Cov}(X_1, X_2) & \cdots & \mathrm{Cov}(X_1, X_p) \\ \mathrm{Cov}(X_2, X_1) & \ddots & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_p, X_1) & \cdots & \cdots & \mathrm{Var}(X_p) \end{pmatrix} = \begin{pmatrix} \sigma^2_{x_1} & \sigma_{x_1 x_2} & \cdots & \sigma_{x_1 x_p} \\ \sigma_{x_2 x_1} & \ddots & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{x_p x_1} & \cdots & \cdots & \sigma^2_{x_p} \end{pmatrix}$$

Figure 9: Définition formelle matrice variance covariance

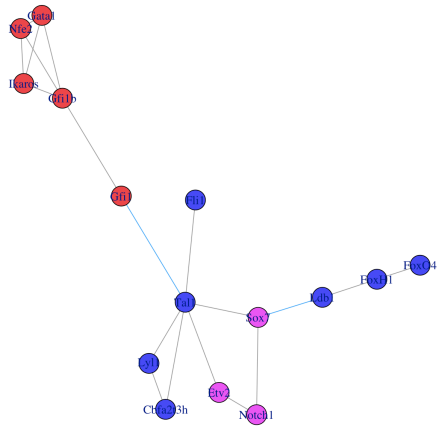

Figure 10: TF partial correlation graph, seuil=0.4, lambda=0.01

En faisant varier la taille du paramètre lambda sur la trace de notre matrice de variance covariance, nous faisions en réalité varier la variance de l'expression de chacun des TF. Plus la valeur de lambda étant élevée, plus la variance l'était également, ainsi lambda influait directement sur l'expression de nos TF: une variance élevée

indiquant alors que les valeurs sont très écartées les unes des autres, et vice versa. La variance étant strictement définie positive, la trace de notre matrice de variance covariance était innévitablement positive tant que nous conservions un lambda positif. Ainsi, un lambda élevé signifiait potentiellement une augmentation du nombre de noeuds sur le graphe, considérés alors comme utiles à la capture de la sémantique du jeu de données pour la modélisation.

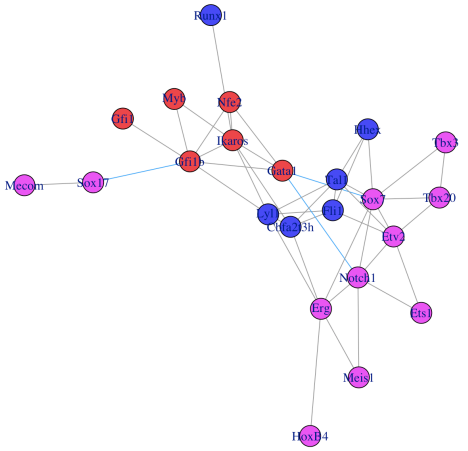

Figure 11: TF partial correlation graph, seuil=0.4, lambda=0.1

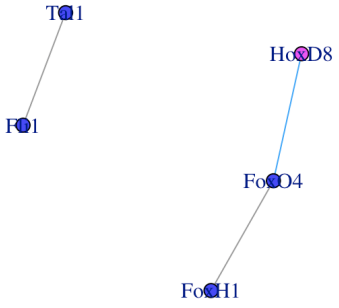

Figure 12: TF partial correlation graph, seuil=0.4, lambda=0.0001

Nous remarquons que, grâce aux corrélations partielles, ces types de graphes semblent capturer les trois groupes de TF que nous avions trouvé en partie 1, contrairement à nos graphes utilisant seulement les corrélations deux à deux qui ne

capturaient pas facilement la totalité des groupes lors de la représentation: ces graphes nous fournissent donc plus d'informations que les précédents.

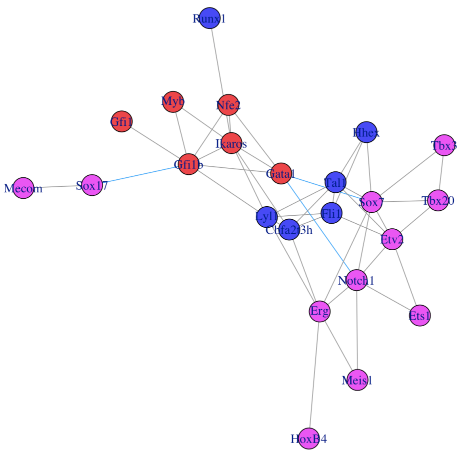

Figure 13: TF partial correlation graph, seuil=0.4, lambda=0.1

Ce graphe nous permet de remarquer que les groupes trouvés par notre méthode experimentale ne diffèrent pas trop du résultat de [**1**]. Ainsi dans les trois groupes, nous pouvons remarquer des similarités pour les clustering de gènes mais également dans les répressions: Gfi1lb, Sox17, Gata1, Sox7, Notch1 sont tous des noeuds dont certaines liaisons sont en réalité des répressions envers les autres TF appartenant à d'autres groupes. Nos résultats, bien que minimes par rapport à la méthode Miic sont en accord avec [**1**].

D'un point de vue biologique, comme précisé précédement, ces graphes pourraient être grandement améliorés: l'expression d'un TF pourrait freiner le taux d'expression d'un autre TF sans que l'inverse soit vrai, ce que nos graphes non dirigés ne capturent pas. C'est probablement une des raisons pour lesquelles nos résultats sont sans doute moins fournis que [**1**]: ici nos repressions sont doubles.

4 Pour aller plus loin

4.1 Utilisation du package miic

Nous avons souhaité pousser l'analyse plus loin en utilisant directement les méthodes de création de graphes du package miic: Multivariate Information based Inductive

Causation [**1**]. Nous permettant alors de comprendre la modélisation de [**1**] comparé à nos deux tentatives de construction de graphes.

Miic est une méthode proposée spécifiquement pour ce problème, qui apprend en découvrant progressivement la contribution indirectes des variables. De cette manière Miic capture les effets directs et indirects entre les variables potentiellement corrélées. Au final, il permet d'obtenir une représentation du réseau en prenant en compte des informations que nous n'avions pas considérée précédement.

En partant d'un graphe complet, la méthode enlève de manière itérative les arrêtes considérées comme superflues. Celles-ci sont considérées comme superflues dès qu'une contribution provenant d'un chemin indirect plus importante est découverte. Les arrêtes restantes étant alors orientées en se basant sur la causalité calculée.

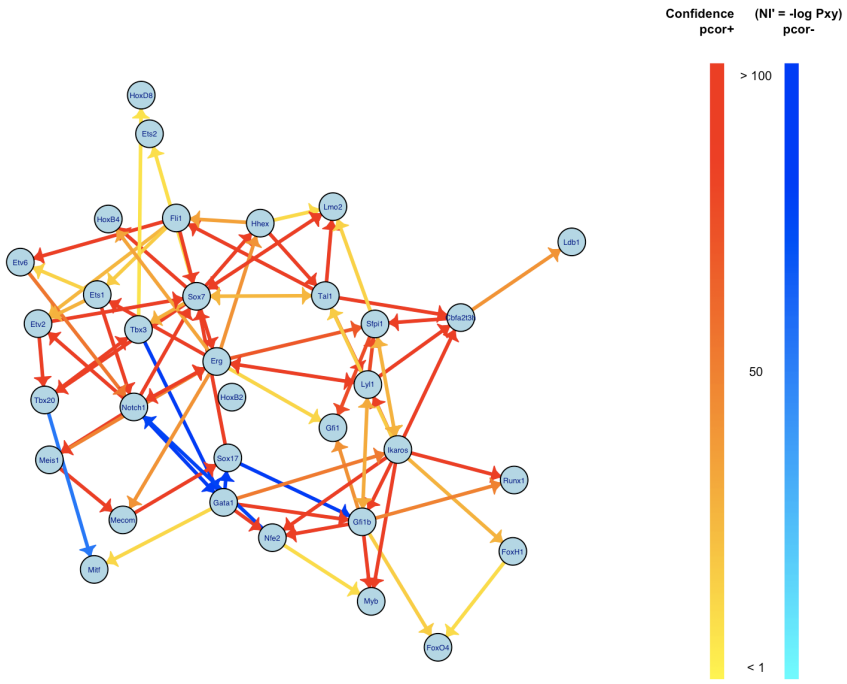

Figure 14: Miic graph: confidenceShuffle = 10, confidenceThreshold = 0.001

Cependant, à partir du moment où l'algorithme coupe une arrête qu'il considère comme superflue alors qu'elle ne l'est pas, l'itérativité de l'algorithme fait que la représentation finale en sera affectée.

5 Conclusion

Dans ce rapport, nous avons eu l'occasion de reconstruire un réseau de gênes et de l'analyser afin de mieux comprendre l'impact de certains facteurs de transcriptions dans l'expression des gènes et donc le développement cellulaire de certaines cellules souches spéecialisées.

Après une brêve analyse des données, et un regroupement des facteurs de transcriptions en trois groupes distincts représentant trois types de cellules souchess dans le vivant, nous nous sommes intéréssés à la construction de deux graphes: l'un prenant en compte les corrélations deux à deux, l'autre prenant en compte les corrélations partielles des facteurs de transcription. Nous avons ensuite comparé ces deux modèles au modèle de [**1**] afin de voir si nos méthodes de construction étaient viables d'un point de vue biologique.

Nous avons également souhaité manipuler la méthode de modélisation de [**1**] afin de comprendre au mieux nos différences de résultats mais également afin d'apprehénder au mieux la suite du cours sur la construction de graphe et de l'information mutuelle.

References

[1] Nadir Sella, Louis Verny,Severine Affeldt, Hervé Isambert *Learning Causal or Non-Causal Graphical Models Using Information Theory.*