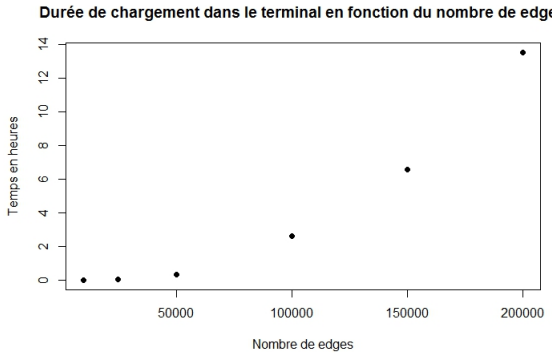
Mike Vadi & Charline Carron Superviseur : Daniel Marbach 09.06.2016

The Facebook of genes : community identification in biological networks

A ce jour, nous avons connaissance de beaucoup de pathways biologiques, nous savons donc quels sont les gènes qui les composent. Cependant, il serait prétentieux de penser qu'un pathway nous ait dit tous ces secrets, il est très probable que nous ignorons encore passablement beaucoup du fonctionnement de tel ou tel pathway. Pour palier à cette ignorance, nous nous sommes demandé s'il y avait des gènes impliqués dans ces pathways qui n'y figurent pas encore, et surtout comment les découvrir ? Pour cela, nous nous sommes inspiré d'une des méthodes qu'utilisent le réseau social Facebook pour proposer à ces utilisateurs de "nouveaux amis potentiels" : Facebook va chercher des informations chez les amis de l'utilisateur pour ensuite essayer de prédire qui pourrait correspondre à ce genre de personnes, et ainsi lui proposer d'inclure le fameux "nouvel ami potentiel" au réseau. Nous allons donc utiliser la même méthode afin de savoir s'il est possible de prédire des gènes pouvant faire partie d'un pathway à partir des gènes qui le composent déjà.

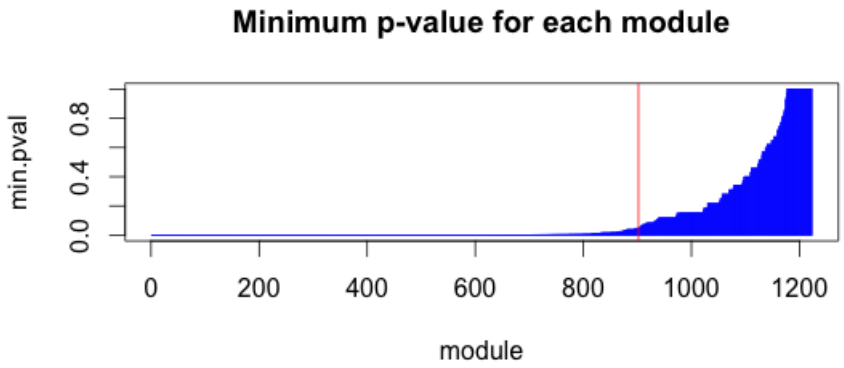
Nous avons utilisé un network composé d'une multitude de gènes d'Homo Sapiens codant surtout pour la synthèse de protéines, liés entre eux par un peu moins de 400'000 liens. Nous avions également à notre disposition 1330 pathways. Le but était donc de chercher si des gènes de notre network pouvaient être condensés en modules, et si ces modules pouvaient faire des overlap (chevauchements) avec les gènes des pathways existants. Nous avons principalement travaillé sur R et utilisé le terminal pour réaliser ce travail.

La première étape que nous devions franchir était celle de savoir combien de edges (nos liens entre les gènes du network) nous voulions garder pour créer les modules, car il était évident que de travailler avec 400'000 edges prendraient trop de temps. nous avons donc réalisé plusieurs essais sur le terminal pour trouver le nombre optimal de edges à choisir :

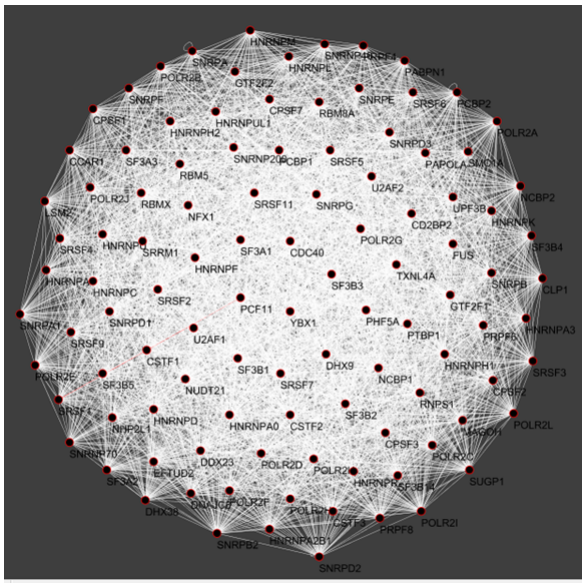


Nous avons remarqué qu'à partir de 200'000 edges, on avait déjà besoin d'environ 14h pour finir de créer les modules, cela ne servaient donc à rien d'essayer avec plus de edges. Pour continuer notre travail, nous avons décidé d'utiliser le network contenant 50'000 edges.

L'étape suivante consistait à présent à comparer nos modules avec les pathways. Pour cela nous avons crée plusieurs matrices : la première nous indiquait grâce aux mentions TRUE/FALSE si les gènes étaient présents dans les modules. La seconde, également une matrice TRUE/FALSE faisaient de même entre modules et pathways. La troisième était une matrice qui nous retournait les p-values correspondant à s'il y avait des overlap entre les modules et tel ou tel pathway. Elles ont été calculées grâce à une loi hypergéométrique puis corrigées grâce à une correction de Benjamini & Hochberg. Nous avons illustré les p-values de cette dernière matrice par le graphe ci-dessous (l'abline rouge correspond au seuil 5% de significativité) :



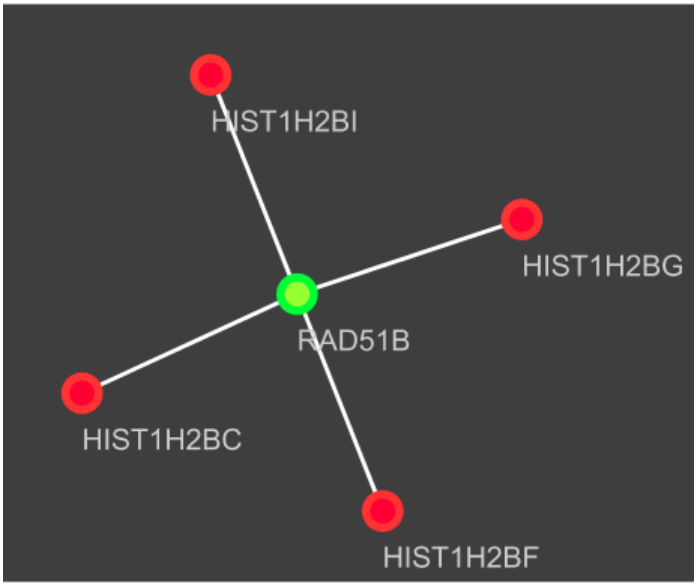
Nous avons donc environ 900 modules significatifs. Ce sont donc des modules dont les gènes ont fait une overlap avec des gènes contenus dans les pathways connus ! Il est intéressant à présent de visualiser ces réseaux afin de voir si l'on trouve de nouveaux gènes pouvant peut-être être inclus aux pathways. Nous avons utilisé le logiciel Cytoscape pour réaliser des visualisations des résultats. En premier, nous nous sommes intéressé au module possédant la p-value la plus significative (Module 1, p-value = $9.15e^{-181}$), le résultat ci-contre a été crée grâce à Cytoscape :



Ce module contenait 106 gènes et nous voyons qu'ils sont tous inclus dans le pathway d'épissage de l'ARNm. Ce résultat est surprenant car nous n'avions pas énormément de chance de

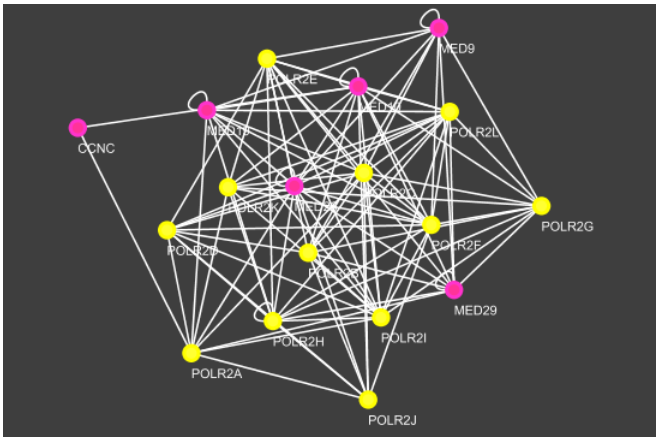
trouver un modèle entièrement dépendant d'un pathway. Malgré tout, il n'illustrait guère le véritale but de notre projet, il fallut donc choisir d'autres modules.

Nous avons choisi de prendre un module ayant une p-value moins signicative, c'est-à-dire le module 7 où la p-value valait $7.56e^{-81}$. Il était composé de 5 gènes, dont 4 qui faisaient des overlap avec le pathway PACKAGING_OF_ TELOMERE_ENDS (en rouge). Le gène est vert est celui qui a été prédit grâce au pathway, il est d'ailleurs intéressant de s'intéresser à la fonction de ces gènes : ceux en rouges sont des gènes histones (HIST1) et le vert (RAD51B)est un gène essentiel à la réparation d'ADN par recombinaison homologue. On peut donc supposer que ce gène a également une place importante dans le pathway concernant les extrémités des télomères.



Nous souhaitions encore visualiser un dernier module contenant plus de gènes que le précédant, nous avons choisi le module 447 ayant une p-value de $3.02e^{-06}$.

Le module 447 possédait 18 gènes dont 12 qui faisaient des overlap avec le pathway VIRAL_MESSENGER_RNA_SYNTHETIS, ces gènes (jaune) étaient tous des gènes impliqués dans la RNA polymerase II. Le gène CCDC (rose) est impliqué quant à lui dans la phosphorylation de l'extrémité C-terminal de la RNA polymerase II. Les 5 derniers gènes appartiennent au complexe MED (Mediator) qui agit comme un coactivateur en se liant à l'extrémité C-terminal de la RNA polymerase II, faisant ainsi un pont entre cette enzyme et les facteurs de transcription.



Ce dernier exemple nous permettra de conclure. Nous avons vu grâce à ce projet qu'il était possible de "prédire" l'implication de certains gènes dans des pathways déja connu. Il serait à présent intéressant d'étudier certains de ces overlap en détail afin de déterminer si les gènes prédits appartiennent effectivement aux pathways de façon significatifs, ou si cela est juste dû au hasard. Néanmoins, nous avons pu voir que cette méthode de recherche pouvait déjà nous donner un bon aperçu de ce que pourrait contenir les pathways de demain.