PROJET EN MODÉLISATION QUALITATIVE
MASTER BIOINFORMATIQUE

Recherche des régulateurs clés de la réponse au nitrate dans les racines d'*Arabidopsis Thaliana* à partir d'un réseau de co-expression

13 novembre 2016

Julie HARDY
Thomas RIQUELME

I. Les enjeux sociétaux

Le nitrate représente la principale source d'azote pour les plantes supérieures. Son absorption par les racines à partir du sol, sa distribution et sa métabolisation dans la plante, sont des étapes essentielles permettant une croissance optimale[1].

En agriculture, la fertilisation des cultures par des engrais azotés est utilisée en vue d'augmenter leur rendement. Ils apportent les éléments nutritifs dont les plantes ont besoin pour se développer et qui peuvent manquer dans les sols trop exploités. Cependant l'administration de ces engrais azotés à des doses massives est responsable d'une pollution importante des sols et des eaux souterraines.

En effet, les nitrates étant très solubles dans l'eau, lorsqu'ils ne sont pas consommés par les plantes, s'infiltrent aisément dans le sol et atteignent progressivement les eaux souterraines. Or celles-ci constituent les principaux réservoirs d'eau potable, et au delà d'un certain seuil (50mg/L selon l'OMS) la concentration en nitrate rend l'eau non potable. Cela constitue un risque pour la santé et l'environnement.

En ce qui concerne les risques pour la santé, les nitrates lorsqu'ils sont ingérés par l'homme, vont être dégradés par une bactérie et se transforment en nitrites. Au-delà d'un certain seuil, ces nitrites peuvent empoisonner le sang en oxydant l'hémoglobine. Le fluide fixe alors mal l'oxygène et engendre des troubles respiratoires. On appelle cet empoisonnement la maladie bleue, ou la méthémoglobinémie. Les nourrissons y sont particulièrement sensibles[2].

En ce qui concerne les risques pour l'environnement, les concentrations massives en nitrates présentes dans l'eau entraînent une eutrophisation des rivières à débit lent, des lacs, des réservoirs et des zones côtières. Cette eutrophisation provoque une prolifération d'algues masquant le fond de la lumière solaire. C'est alors tout l'environnement aquatique qui est modifié, le milieu devenant anoxique, de nombreuses espèces disparaissent, au détriment d'autres[3].

En résumé, l'utilisation massive d'engrais azotés dans l'agriculture entraîne une pollution des sols et de l'eau qui a des conséquences néfastes pour la santé et l'environnement : dangereux pour l'homme si ingéré car les nitrates empoisonnent le sang en oxydant l'hémoglobine et dangereux pour l'environnement en raison des phénomènes d'eutrophisation des rivières, lacs, réservoirs et zones côtières.

Dans ce rapport, une recherche est menée sur les régulateurs clés de la réponse au nitrate dans les racines d'*Arabidopsis Thaliana* à partir d'un réseau de co-expression. Dans un premier temps, les paires de gènes impliquées dans la réponse au nitrate et co-exprimées vont être récupérées à l'aide du logiciel R à partir d'un profil d'expression génique dans différentes expériences de stimulation au nitrate. Puis dans un second temps, un réseau de co-expression sera construit et analysé à l'aide du logiciel Cytoscape pour permettre la sélection des meilleurs gènes candidats régulateurs de la réponse au nitrate dans les racines.

II. Récupération des paires de gènes co-exprimés à l'aide du logiciel R

La première étape consiste en la récupération de profils d'expression génique collectés à partir de 10 expériences de stimulations au nitrate disponibles dans différentes bases de données publiques (cf annexe énoncé) et regroupés dans un fichier "P_EXPR.txt".
Une table contenant la liste des noms des locus des gènes d'*Arabidopsis Thaliana* connus pour répondre à une stimulation à l'azote a aussi été récupérée dans un fichier "P_GOI.txt".
Ces 2 fichiers ont été chargés dans une dataframe dans le logiciel R.

```r
setwd('Documents/M2_semestre9/Qualitative_quantitative_modeling_biology_system /projet_mocell/')
"Lecture des profils d'expression génique dans les différentes expériences"
data=read.table("P_EXPR.txt",header=T)
data
class(data)

# Lecture de la liste des gènes d'intérêt et chargement dans un vecteur
focus<-read.table("P_GOI.txt", header=TRUE)
list_focus <-as.vector(focus[,"id"])
list_focus
```

	id	GSM1054974_Col_0_KCl_2H_R1.CEL.gz	GSM1054975_Col_0_KCl_2H_R2.CEL.gz
1	ATMG00640	5.481737	5.683637
2	ATMG00650	5.614836	5.410114
3	ATMG00660	7.658607	7.390298
4	ATMG00670	6.024494	6.171044

FIGURE 1 – Profils d'expression génique

	id
1	AT1G01190
2	AT1G02310
3	AT1G02340
4	AT1G03080

FIGURE 2 – Liste des gènes d'intérêt

Ensuite, un filtrage est réalisé sur les données d'expression pour récupérer uniquement celles correspondant aux gènes d'intérêt.

```r
# Filtrage des données d'expression correspondant aux gènes d'intérêt
filter_list <- data[,"id"] %in% list_focus
filter_list
data_focus<-data[filter_list,]
data_focus

# Nécessaire de décaler la première colonne pour récupérer les labels des l
rownames(data_focus)

# Les labels que l'on souhaite sont dans la première colonne
data_focus[,1]
rownames(data_focus)=data_focus[,1]
data_focus
data_focus=data_focus[,-1]
data_focus
```

	GSM1054974_Col_0_KCl_2H_R1.CEL.gz	GSM1054975_Col_0_KCl_2H_R2.CEL.gz
AT2G41660	8.725772	8.687922
AT2G41560	6.410440	7.049813
AT1G67810	9.324915	8.905741
AT4G25620	7.960665	7.771613

FIGURE 3 – Profils d'expression des gènes d'intérêt

Puis, les corrélations entre les valeurs d'expressions des gènes sont calculées.

```r
# La fonction t() permet d'inverser les colonnes et lignes - nécessaire pour pouvoir
#calculer les corrélation entre les gènes qui au départ correspondent aux lignes-
data_focus=t(data_focus)
data_focus

# calculer la corrélation entre toutes les colonnes d'une matrice
#(nécessaires de donner en paramètres une dataframe)
result_correlation=cor(data_focus)
result_correlation
```

	AT2G41660	AT2G41560
AT2G41660	1.000000000	0.57747883
AT2G41560	0.577478835	1.00000000
AT1G67810	-0.541406132	-0.72482938
AT4G25620	0.018841127	0.28393712

FIGURE 4 – Résultats des corrélations

Un filtrage sur ces valeurs de corrélation est appliqué en utilisant comme valeur seuil 0.75 pour ne récupérer que les gènes les plus corrélés. La valeur absolue des corrélations a été calculée avant cela pour récupérer les corrélations à la fois positives et négatives supérieures au seuil de 0.75. Enfin, les données sont traitées pour ne récupérer que les paires de gènes co-exprimés (avec une corrélation de leur valeur d'expression supérieure à 0.75), et un type d'interaction est ajouté : "correlation pair" correspondant aux "paires de gènes co-exprimées".

```r
#Utiliser la fonction melt pour convertir une matrice en liste d'interaction
library("reshape2", lib.loc="/autofs/netapp/account/cremi/triquelme/R/x86_64-pc-linux-gnu-library/3.2")
list_correlation=melt(result_correlation)
list_correlation

#On sélectionne uniquement les correlations dont la valeur est supérieur à 0.75
list_correlation=list_correlation[abs(list_correlation["value"])>0.75,]

#Il est nécessaire de supprimer les paires redondantes (on est partie d'une matrice symétrique)
#Création d'une nouvelle colonne Alphabétique pour stipuler si le nome du gene A est ordonné
#selon l'aphabet par rapport au nom du gene B
list_correlation["Alphabétique"]<-as.character(list_correlation[,"Var1"])<as.character(list_correlation[,"Var2"]
list_correlation

#Supression des lignes où l'ordre de classement entre les deux noms de gènes ne suit par
#l'ordre alphabétique (permet de supprimer les doublons)
list_correlation=list_correlation[list_correlation[,4]==TRUE,]
list_correlation

#Supression de la colonne transitoire alphabétique - n'est plus nécessaire-
list_correlation=list_correlation[,-4]

#Supression de la colonne Value
list_correlation=list_correlation[,-3]

#Ajout de la colonne Interaction pour donner le type, ici ce sont les paires dont
#l'expression est corrélée.
list_correlation["type"]="correlation_pair"
list_correlation
```

	Var1	Var2	type
399	AT2G26650	AT2G41660	correlation_pair
682	AT1G80380	AT2G41560	correlation_pair
685	AT1G63940	AT2G41560	correlation_pair
687	AT1G73920	AT2G41560	correlation_pair

FIGURE 5 – Paires de gènes co-exprimés

Dans un second temps, la table des paires de gènes d'*Arabidopsis Thaliana* codant pour des protéines en interaction est chargée (P_PPI.txt). Celle-ci est filtrée pour récupérer uniquement les paires de gènes présents dans la liste d'intérêt.

```
#Lecture de la tables d'interactions Protéine-Proteine de ces gènes
data_PPI=read.table("P_PPI.txt",header=T,sep='\t')
data_PPI

# Filtrage des données PPI correspondant aux gènes d'intérêt
#Premier filtrage sur les gènes de la colonne 'id1'
filter_PPI_id1 <- data_PPI[,"id1"] %in% list_focus
filter_PPI_id1
data_PPI_focus<-data_PPI[filter_PPI_id1,]
data_PPI_focus

#Deuxième filtrage sur les gènes de la colonne 'id2'
filter_PPI_id2 <- data_PPI_focus[,"id2"] %in% list_focus
filter_PPI_id2
data_PPI_focus2<-data_PPI_focus[filter_PPI_id2,]
data_PPI_focus2

#Changement du nom des deux premières colonnes pour qu'elles correspondent dans les deux
#listes
colnames(list_correlation)[1]="id1"
colnames(list_correlation)[2]="id2"
list_correlation

colnames(data_PPI_focus2)[3]="type"
data_PPI_focus2
```

	id1	id2	type
269	AT1G68670	AT2G41730	PPI
3120	AT1G70410	AT3G54900	PPI
4167	AT1G63940	AT2G15620	PPI
4617	AT2G30360	AT5G47100	PPI

FIGURE 6 – Paires de gènes d'intérêt codant pour des protéines en interaction

Enfin, les paires de gènes co-exprimées et les paires de gènes codant pour des protéines en interaction sont formatées de la même façon (gène1, gène2, type d'interaction) et enregistrées dans un fichier au format SIF en vue d'une utilisation dans le logiciel Cytoscape pour construire un réseau de co-expression enrichi avec les données d'interactions Protéine-Protéine.

Ce schéma résume les différentes étapes de l'analyse avec le logiciel R. A chaque étape est indiquée le nombre d'entrées obtenues.

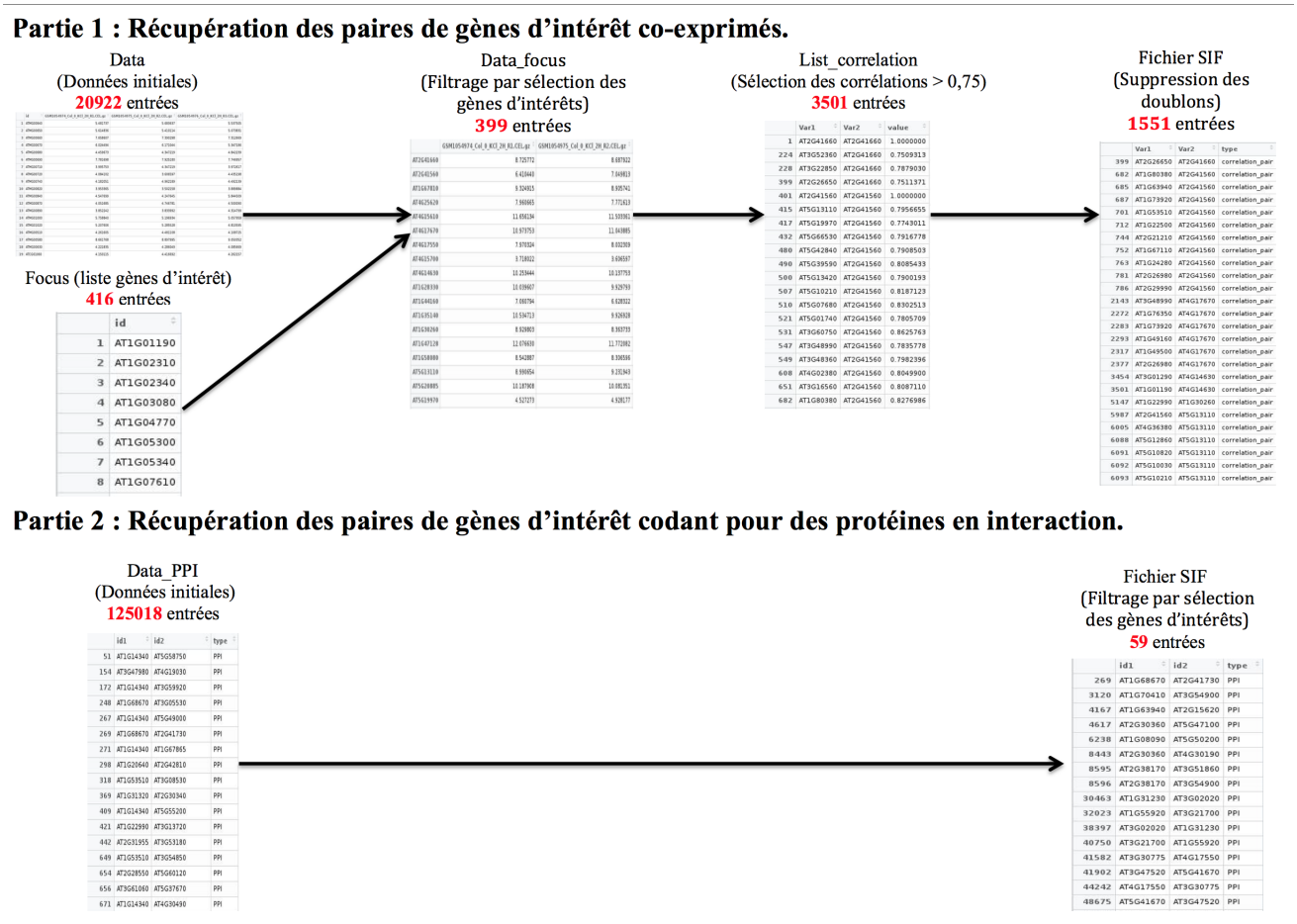

FIGURE 7 – Schéma récapitulatif des résultats obtenus avec le logiciel R

III. Analyse du réseau de co-expression avec le logiciel Cytoscape

III..1 Identification des gènes impliqués dans la réponse au nitrate

Après avoir récupéré les paires de gènes co-exprimées et les paires de gènes codant pour des protéines en interaction grâce au logiciel R, la visualisation du réseau de co-expression est réalisée à l'aide du logiciel Cytoscape.

Le chargement du fichier sif ainsi que du fichier des facteurs de transcription dans cytoscape est réalisé par la commande File → import → network (pour la liste des paires de gènes et type d'intéractions) et File → import → Table (pour la liste des facteurs de transcription).
Un réseau de co-expression est alors construit automatiquement en reliant chaque paire de gènes avec un algorithme qui minimise le croisement d'arêtes (figure 8).

On observe dans ce réseau de co-expression global, un bloc d'interactions principal en haut de la figure 8 où la majorité des intéractions se font et des gènes plus isolés en dessous.
Les gènes qui codent pour des facteurs de transcriptions sont représentés sous forme de triangles rouges grâce à la commande Control Panel→node→shape et Fill Color. Les paires de gènes co-exprimés sont reliées par des arêtes vertes et les paires de gènes codant pour des protéines en interaction sont reliées par des arêtes rouges grâce à la commande la commande Control Panel→edge→shape et Fill Color.

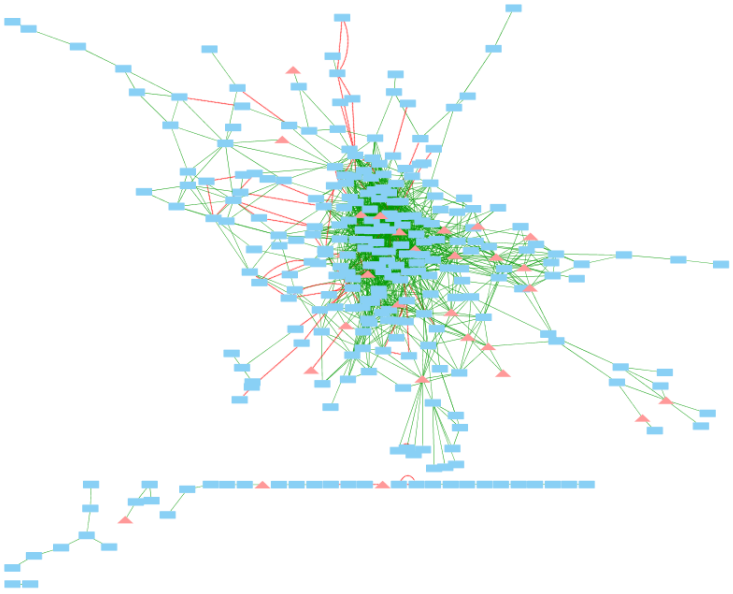

FIGURE 8 – Réseau de co-expression

Afin d'identifier les gènes candidats régulateurs clés de la réponse au nitrate au sein de ce réseau de co-expression, le plugin "BiNGO" a été utilisé sur l'ensemble des gènes du réseau pour identifier ceux impliqués dans la réponse au nitrate.
En effet, le plugin BiNGO permet d'analyser l'implication des gènes dans des processus biologiques répertoriés dans la Gene Ontology (GO) en calculant l'enrichissement, c'est-à dire la surreprésentation des gènes lié à un processus particulier au sein de ce réseau.
Un nouveau réseau d'enrichissement est construit grâce au plugin, correspondant aux relations entre les termes de la GO dans lesquels sont impliqués les gènes du réseau (figure 9). Une représentation hiérarchique est choisie et le layout est changé par la commande suivante Layout→yFilesLayout→hierar

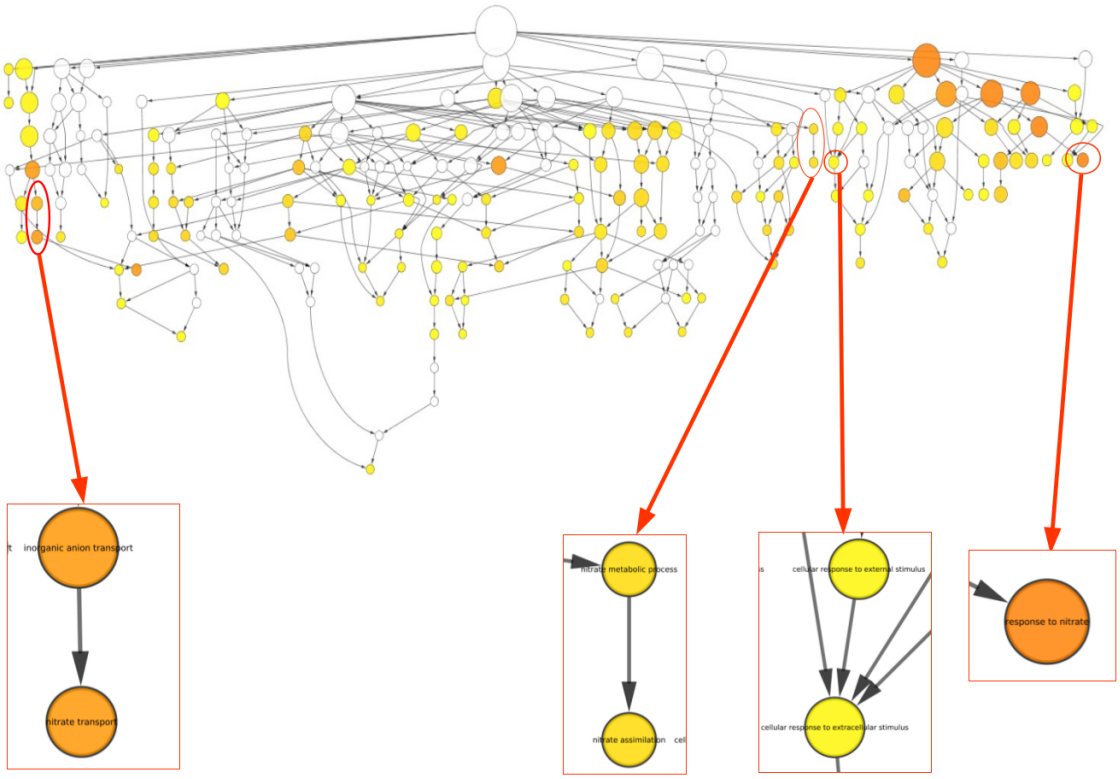

FIGURE 9 – Représentation hiérarchique des intéractions entre les processus biologiques impliqués dans le réseau de co-expression

Le plugin Bingo permet aussi l'obtention d'une table avec l'ensemble des résultats d'enrichissement. Parmi les processus impliqués dans notre réseau, les processus en lien avec la réponse au nitrate ont été sélectionnés :

- réponse au nitrate
- transport du nitrate
- processus métabolique du nitrate
- assimilation du nitrate
- réponse cellulaire à un stimulus externe

Le choix des 4 premiers processus a été simple car ils contenaient le terme "nitrate". Cependant les gènes utilisant ces annotations étaient au nombre de 13 et pour répondre à la consigne de trouver 20 gènes candidats régulateurs de la réponse au nitrate, un cinquième processus a été choisi "réponse cellulaire à un stimulus externe" car dans nos expériences le stimulus externe était la stimulation au nitrate.

De plus, l'enrichissement lié à ces processus est significatif (visible par les couleurs jaunes et oranges de ces nœuds sur le graphe de la figure 9) :
L'enrichissement lié à la réponse au nitrate a une p-value de 2.93^{-13} très significative. Il y a 2.93^{-13} chances que cette sur-représentation des gènes impliqués dans ce processus soit due au hasard. On peut donc en déduire que les gènes impliqués dans la réponse au nitrate sont très représentés dans notre réseau de co-expression. Ce qui est logique vu qu'un premier filtrage a été effectué pour ne garder que les gènes connus pour répondre à une stimulation à l'azote.
Dans une moindre mesure, les enrichissements liés au transport du nitrate, à l'assimilation du nitrate et à la réponse cellulaire à un stimulus externe sont significatifs eux aussi avec des p-values respectives de 3.62^{-8}, 1.03^{-4} et 2.88^{-3}.

Ensuite, les gènes impliqués dans ces processus ont été récupérés grâce à la commande "select nodes". Ils sont répertoriés dans la table ci-dessous (figure 10) et leurs interactions sont représentées sur le réseau de la figure 11.

name	TF_family	__glayCluster
AT5G53460		1
AT1G08090		2
AT1G12110		2
AT1G13300	G2-like	2
AT1G32450		2
AT1G78000		2
AT3G48360		2
AT4G02380		2
AT4G21680		2
AT1G12820		3
AT1G66200		3
AT1G78050		3
AT2G15620		3
AT5G50200		3
AT1G77760		4
AT5G61420	MYB	4
AT1G55920		21
AT4G35090		21
AT1G08100		28

FIGURE 10 – Table des 19 gènes impliqués dans la réponse au nitrate

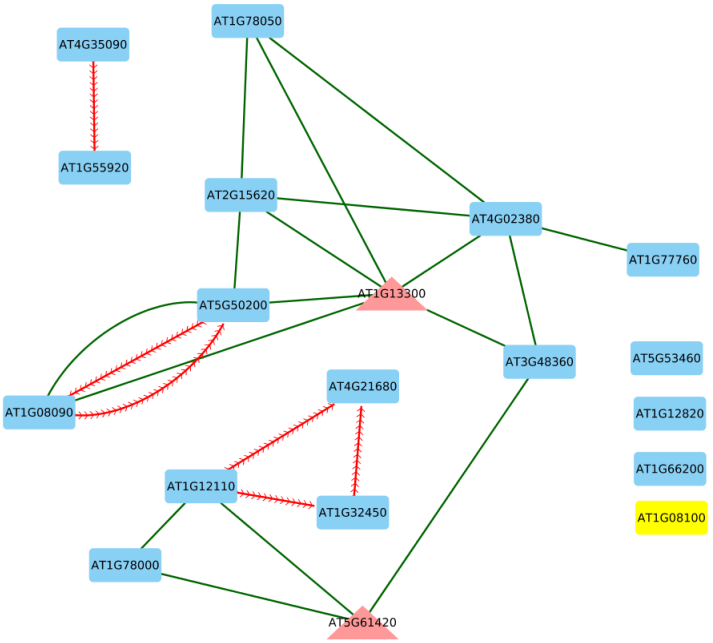

FIGURE 11 – Réseau de co-expression des 19 gènes impliqués dans la réponse au nitrate

On peut supposer que ces gènes sont les meilleurs candidats en tant que "régulateurs clés" de la réponse au nitrate dans les racines d'*Arabidopsis Thaliana*.

On remarque dans la table des gènes candidats (figure 10) que les gènes AT1G13300 et AT5G61420 sont des facteurs de transcriptions appartenant respectivement aux familles de facteurs de transcriptions G2-like et MYB. L'expression du premier (AT1G13300) est corrélée avec l'expression de 6 autres gènes du réseau tandis que l'expression du second (AT5G61420) est corrélée avec l'expression de 3 gènes du réseau dont un en commun(AT3G48360). On peut donc supposer que ces facteurs de transcriptions régulent positivement ou négativement l'expression des autres gènes avec qui ils sont en contact sur le réseau de la figure 11, et même qu'ils régulent ensemble l'expression du gène AT3G48360 avec qui ils sont tous les deux corrélés sur le réseau de la figure 11.

D'autre part, on observe sur la figure 11 que trois gènes codent pour des protéines en interaction (AT4G21680, AT1G12110, AT1G32450). On peut supposer que les protéines pour lesquelles ils codent forment un complexe protéique ou du moins rentrent en contact deux à deux, pas forcément les trois en même temps.

III..2 Identification de clusters

La méthode suivante consiste en l'application d'un algorithme de clustering sur le réseau initial. Après sélection du réseau initial, les commandes suivantes sont appliquées :
Apps →ClusterMaker→Community Cluster(GLay), Cocher : Create New clustered network.

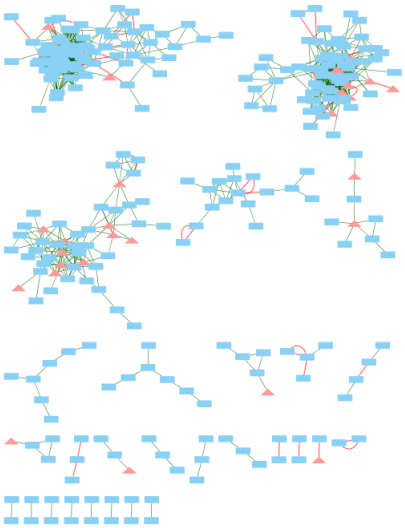

FIGURE 12 – Représentation des différents clusters du réseau

A partir du réseau global, sont définit les différents clusters. On observe que les facteurs de transcription se retrouvent majoritairement dans deux clusters différents.

Sur la représentation des différents clusters, on repère trois clusters importants du fait de leur nombreuses interconnexions avec de multiple gènes. Ils ont donc potentiellement un rôle majeur.

Le cluster n°2 semble être le plus intéressant pour la régulation de la réponse au nitrate. En effet il présente l'enrichissement pour la réponse au nitrate le plus significatif avec une p-value égale à 1.37^{-9}.

Dans une moindre mesure, le cluster n°3 est aussi intéressant et présente un enrichissement pour la réponse au nitrate significatif avec une p-value de 4.65^{-7}.

Néanmoins, l'étude du cluster n°4 indique que les 6 enrichissements les plus significatifs correspondent à des processus biosynthétiques ou métaboliques. Par consequent, ce cluster n'est pas intéressant pour nous car il ne contient pas ou peu de gènes impliqués dans la réponse au nitrate. Ce cluster est plutôt impliqué dans le métabolisme.

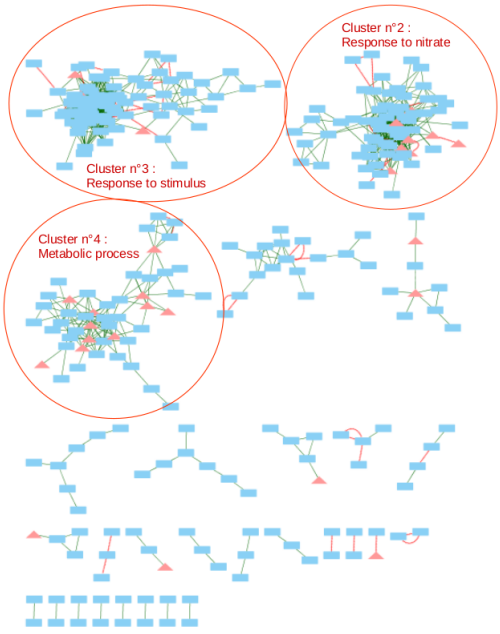

FIGURE 13 – Annotations fonctionnelles de trois clusters

IV. Conclusion

Pour conclure, la sélection des meilleurs candidats pourrait être plus restrictives en prenant en compte uniquement les paramètres associés au mot "nitrate". Cette méthode a pu être testée et a fourni 13 gènes candidats.

Par la suite, pour confirmer par l'expérience que nos 20 gènes sont véritablement des régulateurs clés il pourrait être étudier l'effet de la sur-expression et/ou l'inhibition de ces différents gènes (par exemple avec un KO du gène) sur la plante et en particulier des 2 facteurs de transcription identifiés.

La connaissance des mécanismes des régulateurs permettrait de mieux appréhender l'effet nocif du nitrate. Cette étude parmi toutes les recherches réalisées de nos jours permettrait à long terme de concevoir des programmes d'action efficaces pour résoudre ces problèmes de pollution.

Bibliographie

[1] Céline Mascaux Bertrand Hirel Jean François Morot-Gaudry Mathilde Orsel, Françoise Vedel. Mécanismes moléculaires de l'assimilation de l'azote. In *Colloque national de l'Académie d'Agriculture de France*, 2004.

[2] Delphine Bossy. Engrais, une pollution agricole dangereuse ? `http://www.futura-sciences.com/planete/questions-reponses/pollution-engrais-pollution-agricole-dangereuse-5958/`.

[3] Claude VIDAL Maria PAU VALL. L'azote en agriculture. `http://ec.europa.eu/agriculture/envir/report/fr/nitro_fr/report.htm`. Rapport Eurostat, Commission Européenne.