# Qualitative et quantitative modeling of the cell Projet en modélisation qualitative

(P. Thebault et A. de Daruvar )

# Première partie : recherche de régulateurs clés à partir d’un réseau de co-expression

## Approches globales et utilisation des données publiques

Les approches globales de type « omics » sont de plus en plus utilisées. Elles s’appuient sur un ensemble de technologies en évolution rapide dédiées à l‘analyse des génomes, transcriptomes, protéomes et métabolomes. Elles permettent de produire de nombreux jeux de données à grande échelle : profils d’expression des gènes, interactions protéine-protéine, interactions protéine-DNA. Ils sont disponibles dans les bases de données publiques en libre accès. L’extraction de ces jeux de données permet à n’importe quel laboratoire public ou privé de concevoir et mettre en œuvre des analyses *in silico* pour répondre à ses propres objectifs scientifiques ou industriels.

## Application à la reconstruction de réseau génique

Un réseau génique (GN pour Gene Network) capture les relations entre entités moléculaires qui font partie d'un système biologique. Les GNS sont habituellement représentés sous forme de graphes dont les nœuds représentent des entités moléculaires (telles que des gènes ou des protéines) et les arêtes des relations fonctionnelles entre elles comme par exemple les interactions protéine-protéine, les interactions protéine-ADN (ex : facteurs de transcription) ou encore les relations de co-expression. L’intégration de différents types de données permet une meilleure compréhension de la structure et de la régulation du réseau. Au sein de tels réseaux les nœuds de type « Hub » jouent potentiellement un rôle majeur.

## Description de la problématique biologique

Vous travaillez dans un laboratoire de biologie végétale qui s’intéresse à la régulation de la réponse au nitrate dans les racines de plantes. L’enjeu de ces travaux dépasse celui de la connaissance fondamentale des mécanismes de signalisation moléculaire. En effet, la fertilisation des cultures par des engrais azotés constitue un enjeu sociétal majeur.

## Jeu de données

Vous disposez d’un ensemble de fichiers textes dont vous explorerez la structure et qu’il vous appartiendra de manipuler avec le logiciel R. Ces fichiers sont les suivants :

* **P\_EXPR.txt**: chaque ligne de cette table donne pour un gène *d’Arabidopsis Thaliana* (AT) les valeurs d’expression observées dans 146 mesures. Ces mesures ont été collectées à partir de 10 expériences disponibles dans différentes bases de données (cf. Annexe). Les expériences ont toutes utilisées le microarray Affymetrix ATH1. La première colonne de la table donne le nom des locus des gènes. Les données ont déjà été nettoyées et normalisées par la méthode Robust Multiarray Analysis (RMA).
* **P\_GOI.txt**: cette table donne la liste des noms des locus des gènes d’AT connus pour répondre à une stimulation à l’azote. **On limitera l’étude à cette liste (GOI = Gene Of Interest)**. La table contient une seule colonne « id ».
* **P\_PPI.txt**: cette table donne la liste de paires de gènes d’AT codant pour des protéines en interaction. La table contient 3 colonnes « id1 », « id2 » et « type ». Les colonnes « id1 » et « id2 » contiennent les identifiants des locus des gènes et la colonne « type » contient la valeur « PPI ».
* **P\_TF\_LABEL.txt**: cette table donne une liste des facteurs de transcription d’AT. La table contient 3 colonnes « id », « TF\_family » et « gen\_type ». La colonne « id » contient les identifiants des locus des gènes, la colonne « TF\_family » donne le nom de la famille de facteur de transcription à laquelle appartient ce gène et la colonne « gen\_type » contient la valeur « TF ».

*Ces fichiers ont été préparés à partir des données aimablement mises à disposition par Tomás C. Moyano, Elena A. Vidal, Orlando Contreras-López et Rodrigo A. Gutiérrez du « Plant Systems Biology Lab » de l’université pontificale catholique du Chili. Les données proviennent du laboratoire et de différentes bases de données publiques (cf Annexe).*

## Votre mission

A partir du jeu de données mis à votre disposition, vous utiliserez le logiciel R (RStudio) pour construire un réseau qui combine les informations disponibles sous forme de graphe. Vous commencerez par construire un réseau de co-expression en ne prenant en compte que les gènes d’intérêt (cf jeu de données). Vous enrichirez ce réseau avec les autres relations connues entre ces gènes (cf jeu de données) : attention, l’étude avec cytoscape ne sera pas réalisable si vous ne ciblez pas l’analyse sur les gènes d’intérêts). Vous explorerez ce graphe à l’aide du logiciel Cytoscape pour proposer la vingtaine de gènes qui vous semblent être les meilleurs candidats en tant que « régulateur clé » de la réponse au nitrate dans les racines *d’AT*. Parmi ces gènes vous identifierez les facteurs de transcription. Par ailleurs, à partir de ce réseau, vous construirez des clusters et proposerez des annotations fonctionnelles pour au moins 2 d’entre eux.

## Votre travail

A l’issue de ce projet, vous devrez présenter votre travail aux enseignants et notamment :

* expliquer les enjeux sociétaux liés à la problématique biologique et présenter la démarche utilisée.
* présenter les données utilisées et expliquer pas à pas votre analyse en indiquant précisément les commandes « R » ou « cytoscape » que vous aurez utilisé.
* décrire les résultats obtenus en vous appuyant sur des chiffres, des figures (qui pourront être des copies d’écran).
* Faire une analyse critique de vos résultats et formuler des propositions sur la façon de tester les résultats/hypothèses que vous aurez obtenus.

*Nb : Vous veillerez à citer les sources que vous aurez utilisées  (éviter impérativement toute forme de plagiat).*

# Annexes

## Origine des données de transcriptome

|  |  |  |
| --- | --- | --- |
| **Publication** | **Titre de l’expérience** | **Identifiant (base de données)** |
| Gutierrez (2007) | Transcription profiling by array of Arabidopsis grown in nutrient solutions with various concentrations of nitrate and sucrose. | E-MEXP-828 (ArrayExpress) |
| Krouk (2010) | High resolution NO3– response of Arabidopsis roots | GSE20044 (GEO) |
| Ju (2009) | Expression data of 10-day-old wild-type and *chl1-5* plants exposed to 25 mM nitrate for 0 or 0.5 h | GSE9148 (GEO) |
| Ruffel (2011) | A systemic view of coordinated root responses to NO3– heterogeneous environment in Arabidopsis | GSE22966 (GEO) |
| Patterson (2010) | Comparison of root transcriptomes in *Arabidopsis thaliana* plants supplied with different forms of inorganic nitrogen | GSE29589 (GEO) |
| Vidal (2013) | Root nitrate response of Ws plants and *afb3-1* mutant plants. | GSE35544 (GEO) |
| Alvarez (2014) | Root nitrate response of Col-0 plants and *tga1/tga4* mutant plants | GSE43011 (GEO) |
| Wang (2003) | Treatment of Arabidopsis with low concentration of nitrate. | Exp479 (iPlant Collaborative) |
| Wang (2004) | WT vs. *Nitrate Reductase* null mutant high nitrate concentration treatment. | Exp480 (iPlant Collaborative) |
| Wang (2007) | Arabidopsis treated with nitrite and nitrate | Exp481 (iPlant Collaborative) |

## Autres sources de données

La liste des gènes d’intérêt est disponible sous forme du « Supplementary Material » de la publication suivante : Canales J, Moyano TC, Villarroel E, Gutiérrez RA (2014) Systems analysis of transcriptome data provides new hypotheses about Arabidopsis root response to nitrate treat- ments. Front Plant Sci 5:22.

Les données d’interaction protéine-protéine ont été compilées à partir des bases de données suivantes : PLAPID, ATPID, ATPIN et PAIR.

La table contenant la liste des facteurs de transcription *d’Arabidopsis Thaliana* avec leur famille a été compilée par le laboratoire Plant Systems Biology Lab de l’université pontificale catholique du Chili.