

# Qualitative et quantitative modeling of the cell

## Projet en modélisation qualitative

(P. Thebault et A. de Daruvar : octobre 2016)

## 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 microbienne qui s'intéresse à la résistance des bactéries aux antibiotiques. Grâce à la découverte des antibiotiques, le nombre de décès associé à des maladies infectieuses a nettement diminué au cours du siècle dernier. Cependant, une utilisation intense de ces molécules a impacté l'évolution des bactéries en conduisant à l'apparition de résistances à ces médicaments. Ces résistances aujourd'hui préoccupantes constituent un enjeu sociétal majeur auquel il est nécessaire de s'intéresser pour améliorer les traitements.

### Jeu de données

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

- **P\_EXPR\_antibios.txt** : chaque ligne de cette table donne pour un gène d'*Escherichia coli* MG1655 (EC) les valeurs d'expression observées dans 20 mesures. Ces mesures ont été collectées à partir de 6 expériences

disponibles dans la base de données « GEO database ». Les expériences ont toutes utilisées le microarray Affymetrix *E. coli* Genome 2.0 Array. La première colonne de la table donne le nom des locus des gènes. Les colonnes suivantes correspondent aux résultats d'expression des gènes après un traitement à H2O2, ampicilline, gentamicine, kanamycin, norflacin et aucun traitement. 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'EC connus pour être impliqués dans la résistance aux antibiotiques. La table contient une seule colonne « id ».
- **GOI\_FUR\_1.txt et GOI\_FUR\_2.txt**<sup>1</sup> : ces deux tables sont à exploiter pour identifier des listes de gènes impliqués dans l'utilisation du fer chez EC (des liens entre la résistance aux antibiotiques et l'utilisation du fer ayant été suggérés). Les deux tables contiennent 5 colonnes. Les colonnes « bnum » et « Gene » contiennent les identifiants des gènes et les colonnes « WT\_FeCl2 » et « Dfur\_FeCl2 » les expressions de ces gènes pour une cellule sauvage et un mutant du gène FUR (impliqué dans la régulation des gènes liés au fer). La valeur «  $\log_2(\text{fold\_change})_{\text{Dfur/wt}}$  » contient la valeur de sur ou sous expression des gènes calculée à partir des deux colonnes précédentes. On s'intéressera aux gènes dont **une sous et/ou sur expression** d'un facteur 2 a été observée lors de la comparaison du sauvage/mutant.
- On limitera l'étude à l'union des 3 listes obtenues avec les fichiers (P\_GOI.txt, GOI\_FUR\_1.txt et GOI\_FUR\_2.txt) (GOI = Gene Of Interest). La table contient une seule colonne « id ».
- **sRNA\_list\_coli.txt** : cette table donne la liste des petits ARNs non codant de EC qui modulent la traduction des ARNm. Cette régulation post-transcriptionnelle s'effectue grâce à une interaction entre les deux ARNs. La table contient 2 colonnes « id » et « gen\_type ». La colonne « id » contient le nom du petit ARN et la colonne « gen\_type » contient la valeur « sRNA ».
- **sRNA\_interaction.txt** : cette table donne une liste des cibles (gènes) des petits ARNs. La table contient 3 colonnes « ID », « CIBLES » et « INTERACTION ».
- **P\_TF\_LABEL\_COLI.txt** : cette table donne la liste des facteurs de transcription de EC (59 gènes). La table contient 2 colonnes « id » et « gen\_type ». La colonne « id » contient les identifiants des locus des gènes et la colonne « gen\_type » contient la valeur « TF ».
- **gene\_association-2.ecocyc** : ce fichier contient l'annotation des gènes de ECcoli avec les termes de la GO (gene ontology). Il sera utilisé par le plugin BINGO (le génome de coli n'est pas proposé par défaut).

### 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. ex exploitant les 3 fichiers du jeu de données). Vous enrichirez ce réseau avec le réseau de régulation centré sur les

---

<sup>1</sup> Données issues du supplementary data de « Seo SW, Kim D, Latif H, O'Brien EJ, Szubin R, Palsson BO. Deciphering Fur transcriptional regulatory network highlights its complex role beyond iron metabolism in Escherichia coli. Nat Commun. 2014 Sep 15;5:4910.»

petits ARNs non codant. Vous explorerez ce graphe à l'aide du logiciel Cytoscape pour proposer des gènes qui vous semblent être les meilleurs candidats en tant que « gène clé » de la résistance aux antibiotiques. Parmi ces gènes vous identifierez les facteurs de transcription et ceux ciblés par des petits ARNs. Ces informations vous aideront à définir les gènes dont l'expression pourrait être régulée (par un FT ou sRNA). 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

Vous rédigerez votre rapport en français ou en anglais sous forme d'une mini publication scientifique de 5 pages maximum (police Arial 12, interligne 1,5) hors figures et références, qui sera composée des sections suivantes :

- **Introduction** : vous expliquerez les enjeux sociétaux liés à la problématique biologique et présenterez la démarche.
- **Matériel et méthode** : vous présenterez les données utilisées et expliquerez pas à pas votre démarche en indiquant précisément les commandes « R » ou « cytoscape » de façon à ce que le lecteur puisse reproduire vos résultats.
- **Résultats** : vous décrirez les résultats obtenus en vous appuyant sur des chiffres, des figures (qui pourront être des copies d'écran). Chaque tableau ou figure devra comporter un titre et une légende.
- **Discussion** : vous ferez une analyse critique de vos résultats, vous formulerez des propositions sur la façon de tester les résultats/hypothèses que vous aurez obtenus.

## Origine des données de transcriptome

Publication	Titre de l'expérience	Identifiant (base de données)
Dwyera (20014)	Antibiotics induce redox-related physiological alterations as part of their lethality	GDS5160 (GEO)

### Autres sources de données

Les listes des gènes d'intérêt est disponible sous forme du « Supplementary Material » des publication suivante :

- Seo SW, Kim D, Latif H, O'Brien EJ, Szubin R, Palsson BO. Deciphering Fur transcriptional regulatory network highlights its complex role beyond iron metabolism in *Escherichia coli*. Nat Commun. 2014 Sep 15;5:4910

- Dwyer DJ, Belenky PA, Yang JH, MacDonald IC, Martell JD, Takahashi N, Chan CT, Lobritz MA, Braff D, Schwarz EG, Ye JD, Pati M, Vercruysse M, Ralifo PS, Allison KR, Khalil AS, Ting AY, Walker GC, Collins JJ. Antibiotics induce redox-related physiological alterations as part of their lethality. Proc Natl Acad Sci U S A. 2014 May 20;111(20)

Les données de petits ARNs et de leurs cibles ont été compilées à partir des bases de données suivantes : sRNATarbase3.0.

La table contenant la liste des facteurs de transcription d'*Escherichia coli* a été compilée à partir de la base de données Ecocyc.