Qualitative et quantitative modeling of the cell
Projet en modélisation qualitative

Elodie Darbo, IR Université de Bordeaux et Institut Bergonié
edarbo <elodie.darbo@u-bordeaux.fr

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 recherche en cancérologie qui s'intéresse à la classification des patientes atteintes de cancer du sein en fonction du sous-type tumoral dont elles souffrent. L'enjeu de ces travaux est de permettre aux cliniciens at aux médecins de prescrire la thérapie la plus efficace aux patientes. En effet, il existe cinq sous-types de cancer du sein, dont les mécanismes d'oncogenèse diffèrent, qui ne répondront qu'à la thérapie appropriée. Le cancer du sein est le type le plus fréquent chez les femmes (1% des cancers du sein sont masculins et souvent de mauvais pronostique) et le deuxième cancer de plus fréquent de façon globale ; 58 459 nouveaux cas ont été dépistés en France métropolitaine en 2018 et 12146 décès ont été déclarés [source : Institut National du Cancer : INCa : https://www.e-cancer.fr/]. Il s'agit donc d'un enjeu sanitaire 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 :

- **TCGA_rsem_norm_Nature2012_RNAseq.tab**: chaque ligne de cette table donne pour un gène humain les valeurs d'expression observées dans 825 patientes atteintes de cancer du sein. Ces mesures ont été collectées à partir de la base de données publique The Cancer Genome Atlas (TCGA) [source : https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga]. Le transcriptome des tumeurs a été séquencé avec la technologie Illumina HiSeq. La première colonne de la table donne le nom des locus des gènes. Les données ont déjà été nettoyées et normalisées par la méthode RNA-Seq by Expectation-Maximization (RSEM) [source :Li et Dewey 2011 BMC Bioinformatics].
- **PAM50_geneIDs.txt** : cette table donne la liste des noms des gènes humains de la signature PAM50 contenant 50 gènes dont l'expression permet de prédire à quel sous-type appartient chaque tumeur [source : Wallden et al. 2015 BMC Med Genomics]. **On utilisera cette liste pour visualiser ces gènes sur le réseau**. La table contient une seule colonne « id ».
- **GOI_ids.txt** : cette table donne la liste des noms des gènes humains sur lesquels l'analyse se focalisera. Cette liste a été construite à partir de la liste PAM50 et les gènes reliés.
- **Human_STRING_PPI.tab** : cette table donne la liste de paires de gènes humains codant pour des protéines en interaction (source base de données STRING : https://string-db.org). La table contient 3 colonnes « id1 », « id2 » et « type ». Les colonnes « id1 » et « id2 » contiennent les identifiants des gènes et la colonne « type » contient la valeur « PPI ».
- **Human_TF_LABEL.tab**: cette table donne une liste des facteurs de transcription de l'humain. La table contient 3 colonnes « id », « `TF_family` » et « `gen_type` ». La colonne « id » contient les identifiants 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 ».

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 identifier les fonctions biologiques dans lesquelles sont impliquées les gènes de la PAM50. Pour cela, à partir de ce réseau, vous construirez des clusters et proposerez des annotations fonctionnelles pour au moins 2 d'entre eux.
Parmi les fonctions biologiques impliquées dans le cancer du sein , vous vous appuierez sur l'article fournit sur moodle:

Otto T, Sicinski P. Cell cycle proteins as promising targets in cancer therapy. Nat Rev Cancer. 2017 Jan 27;17(2):93-115

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).