Master 2 bioinformatique
Projet Python (1)

17 septembre 2025

1 Préambule : Contexte Biologique

La recherche exploratoire se caractérise par l'absence d'hypothèse que l'on cherche à valider. Au contraire, en recherche exploratoire, on cherche à identifier de nouvelles potentielles hypothèses. C'est par exemple ce qu'on fait fréquemment lorsque qu'on réalise des analyses -omiques à haut débit entre une condition contrôle et une condition qu'on cherche à caractériser. Avec ces analyses -omiques haut débit, on va (i) mettre en évidence tout un ensemble de gènes ou protéines liés à la condition qu'on cherche à caractériser ; (ii) analyser les processus biologiques liés à cet ensemble de gènes/protéines ; (iii) on en déduira l'hypothèse que les processus biologiques identifiés sont clés dans la condition qu'on cherche à caractériser.
Prenons l'exemple d'une analyse transcriptomique (type NGS), qui compare le transcriptome du tissu adipeux chez des personnes non malades et chez des personnes souffrant d'un syndrome métabolique. L'analyse des données nous permettra d'identifier des gènes différentiellement exprimés (item i). Ensuite, on pourra regarder les annotations de ces gènes/protéines pour identifier dans quel processus cellulaire X,Y,Z ils/elles interviennent (item ii). Enfin, on en concluera que les processus X,Y,Z sont dérégulés chez les personnes souffrant de syndrôme métabolique (item iii).

Dans ce sujet, on s'intéresse plus particulièrement à l'étape (ii), qui consiste, à partir d'une ensemble de gènes/protéines, à identifier quels sont les processus cellulaires clés reliés. Deux grandes familles de méthodes existent pour faire cela.

- Les méthodes ORA pour *Over Representation Analysis*, qui se basent sur la fréquence d'une annotation dans notre ensemble, comparée à la fréquence si on prenait au hasard un groupe de protéine [1]. Par exemple, si l'annotation `DNA damage checkpoint` est retrouvée pour 25% des gènes/protéines du groupe, alors qu'au hasard on s'attend à trouver autour de 10%, il y a sur-représentation de l'annotation `DNA damage checkpoint`. D'un point de vue statistique, les calculs de sur-représentations sont généralement basés sur la loi hypergéométrique, et on les corrige systématiquement pour les tests multiples. De très nombreux outils bioinformatiques existent pour calculer ces enrichissements.
- D'autres méthodes sont basées sur le principe GSEA pour *Gene-Set Enrichment Analysis*. Dans une analyse de type GSEA, on va classer tout les gènes, qu'ils soient différentiels ou non, dans l'ordre du plus sur-exprimé au plus sous-exprimé (ie. du fold-change le plus grand

1. un groupe au hasard ayant exactement la même taille, car cela influe sur le résultat et notamment sur le calcul de la p-value

au plus petit). Ensuite, pour une annotation donnée, on va regarder si les gènes/protéines qui portent cette annotation sont en moyenne situés plutôt au début de notre échelle (l'annotation est pertinente et le processus est sur-exprimé), plutôt au mileu de notre échelle (l'annotation n'est pas pertinente), ou plutôt à la fin de notre échelle (l'annotation est pertinente et le processus est sous-exprimé). De très nombreux outils bio-informatiques utilisent cette technique, et on la dit plus robuste car on analyse tous les gènes et pas seulement les gènes différentiellement exprimés : ainsi, il n'y a pas d'effet seuil autour du seuil de 5% utilisé classiquement pour dire si un gène est différentiellement exprimé ou non.

Ces méthodes existent, et l'objet de ce projet n'est pas de vous demander de les ré-implémenter. Cela dit, on va vous demander d'en implémenter une autre :-)

Les deux approches listées ci-dessous ont cependant un problème : elles renvoient généralement beaucoup (trop) d'annotations pertinente. Cela vient du fait que les annotations des gènes sont très nombreuses, et organisées dans une structure d'ontologie.

Les ontologies sont une très bonne structure de données pour représenter les connaissance que l'on a sur la fonction des gènes [2], car elles permettent de hiérarchiser les annotations et faire des liens entre elles. Par exemple, si une protéine est annotée "positive regulation of transcription", alors, même si ce n'est pas écrit, elle est aussi annotée "regulation of transcription", parce que "positive regulation of transcription" est un enfant de "regulation of transcription".
Cette organisation hierarchisée des connaissances est très souple et permet de représenter la connaissance avec la granularité dont on dispose (on ne sait pas toujours si le régulation est positive ou négative, donc c'est utile de pouvoir dire "régulation positive", "régulation négative" ou juste "régulation"...), mais elle permet aussi de faire des ponts entre différents processus qui participent à une même grande fonction. Par contre, cette organisation des connaissances induit quelques désagréments : (i) il y a beaucoup d'annotations ; (ii) quand on ajoute tous les ancêtres aux annotations d'une protéine, les protéines finissent par avoir énormément d'annotations ; (iii) on ajoutera, pour les plus matheux, que statistiquement ces annotations ne sont pas indépendantes les unes des autres.

La plus célèbre ontologie qui annote la fonction des gènes est connue sous le nom de `Gene Ontology` (acronyme GO) : vous pouvez la visiter le site `https://geneontology.org/`. Dès la page d'accueil, on vous proposera de faire une analyse d'enrichissement :-) Allez explorer la page d'accueil, regaldez un peu les onglets proposés, et essayez par exemple de répondre à la question suivante : quels sont tous les ancêtres (ou termes parents) du terme "G2/M transition of mitotic cell cycle". [3].

2 Objectif

Caractériser un ensemble de protéines en fonction des annotations des protéines qui le composent est un problème difficile, sur lequel la communauté scientifique travaille depuis au moins deux décennies et qui demeure une question ouverte.

Dans ce sujet, on vous demande de tester une nouvelle méthode à l'aide d'un algorithme auquel

2. et sur les connaissances en biologie en général, vous pourrez discuter de cela dans l'UE DEL
3. Au cas où, son identifiant est GO :0000086

nous avons pensé, mais que nous n'avons encore jamais testé. Cet algorithme a la particularité de se baser uniquement sur 3 notions :

1. le contenu d'information des annotations (IC pour *Information Content*) : dans l'idée, on aimerait bien que les processus cellulaires clés reliés à notre ensemble de gènes soient assez précis (ie. et ne soient pas trop vagues).
2. la couverture : dans l'idée, on aimerait que les processus cellulaires clés que l'on choisit représentent une grande partie des gènes/protéines de l'ensemble.
3. les relations parents - enfants entre les annotations : dans l'idée, une fois qu'on a dit "response to stimulus", cela n'a pas vraiment de sens de sélectionner ensuite "positive response to stimulus", car cet aspect était déjà traité par l'annotation précédente.

Vous êtes donc dans une situation de projet typique : nous n'avons pas tout testé au préalable, et nous allons devoir travailler ensemble pour aboutir à une solution. Soyez force de proposition, et soyez prompts à nous remonter vos points de blocage !

Comme c'est une question scientifique ouverte et une nouvelle méthode que nous vous demandons d'implémenter, on attend de vous :

— que vous vérifier bien la correction de votre méthode :est-ce que la méthode respecte les spécifications qu'on vous a données ;
— que vous évaluiez vos résultats en les comparant aux solutions existantes.

3 Déroulement

La méthode à laquelle nous avons pensé, basée uniquement sur le contenu d'information, la couverture, et la notion de parent-enfant dans les annotations est décrite brièvement dans le document PDF çi-joint. Prenez-en connaissance.

La première chose que nous attendons de vous est que vous nous remontiez ce dont vous avez besoin pour mener votre projet à bien. Nous savons déjà que vous aurez certainement besoin de la librairie python `goatools`, pour laquelle vous trouverez de la documentation ici : `https://github.com/tanghaibao/goatools`. Pour le reste, ne bloquez pas si il vous manque quelque chose et que vous nous avez envoyé un email : si vous ne pouvez pas travailler sur une partie, travaillez sur une autre le temps que nous vous répondions. C'est une bonne habitude à prendre et cela vous forcera à séquencer votre projet (ie. le découper en work packages assez autonomes).

Pour nous contacter, merci d'envoyer un email systématiquement à :

— `emmanuelle.becker@univ-rennes.fr` et
— `olivier.dameron@univ-rennes.fr` et
— `marine.jacquier@univ-rennes.fr`.

Nous vous répondrons dès que possible. Dans tous vos emails, vous préciserez dans l'objet `M2 BI - projet annotations`, et vous indiquerez quels sont les membres du binôme.

4 Calendrier

Le rendu du projet est un code + un rapport (environ 4 pages figures incluses) qui devra :
— présenter votre projet et comment vous avez structuré le code
— présenter (et parfois justifier) les choix techniques que vous aurez faits
— que la méthode fait ce qu'on vous a demandé (validation)
— comparer la méthodes aux autres méthodes (évaluation).

Vous devrez transmettre votre rapport à deux personnes du groupe classe (désignées ultérieurement) au plus tard le 2 octobre. Les deux personnes devront vous faire un retour sur votre rapport pour le 6 octobre (en 1 page max par personne). Vous aurez ensuite jusqu'au 8 octobre pour prendre en compte les retours (ou pas si vous ne les jugez pas pertinents), et soumettre les deux évaluations par vos camarades que vous avez reçues + votre rapport final.

Vous serez notés à la fois sur la base de votre rapport, mais aussi sur la base de ce que vous avez fait comme retour sur le travail des autres.