M2 Bioinformatique et Biologie des systèmes
Université de Toulouse - Paul Sabatier

Gestion de données non structurées et applications post-génomiques

Rapport de Projet NeoBlastSets

Analyse d'Enrichissement Fonctionnel

Issa KERIMA-KHALIL

Hawa BALDE

2025-2026

I. Introduction

L'étude des ensembles de gènes est un enjeu central en biologie moléculaire et en génomique, car elle permet de relier les données expérimentales à des fonctions biologiques connues et d'identifier des voies ou processus enrichis dans un sous-ensemble de gènes d'intérêt. L'analyse d'enrichissement, également appelée sur-représentation statistique, est fréquemment utilisée pour interpréter les résultats d'expériences à grande échelle telles que les analyses de transcriptome ou de protéome (Huang et al., 2009). Ces méthodes permettent de déterminer si un ensemble de gènes est significativement associé à certaines catégories biologiques ou fonctionnelles, par exemple des GOTerms, des pathways métaboliques ou des domaines protéiques.

Dans ce projet, nous avons construit une base de données intégrée sous Neo4j, incluant des informations sur les gènes, leurs synonymes (alias), leurs fonctions biologiques (GOTerms, InterPro domains, pathways), les références bibliographiques associées (PubMed), ainsi que des scores d'association entre paires de gènes ou protéines basés sur la coexpression, l'interaction protéine-protéine et les liens phylogénomiques. Cette intégration permet de représenter les gènes et leurs relations sous forme de graphe, facilitant l'interrogation et la recherche d'enrichissement de manière structurée et efficace.

L'objectif principal de ce travail est d'analyser un ensemble de gènes fourni en identifiant les catégories biologiques les plus pertinentes au moyen d'approches d'enrichissement fonctionnel. Cette analyse nécessite l'utilisation d'une mesure fiable pour quantifier la similarité entre un ensemble de gènes query et des ensembles cibles de référence. Afin d'assurer la robustesse de cette étape, nous avons d'abord étendu un script existant, initialement fondé sur le test binomial, pour intégrer trois nouvelles mesures : la loi hypergéométrique (test exact de Fisher), le χ^2 d'indépendance et une mesure naïve de coverage. Ces mesures reposent sur des hypothèses statistiques différentes et présentent des sensibilités variables selon la taille des ensembles et la proportion de signal biologique.

Une comparaison empirique de ces quatre approches a ensuite été conduite à l'aide de requêtes synthétiques dérivées des ensembles cibles de référence. Ce benchmarking méthodologique vise à déterminer quelle mesure fournit les résultats les plus stables et les plus pertinents, afin de l'utiliser pour l'analyse biologique finale. Le rapport présente ainsi l'intégration des données, les modifications apportées au script, l'évaluation comparative des mesures, puis l'analyse détaillée de l'ensemble de gènes étudié et l'interprétation biologique des enrichissements obtenus.

II. Intégration et préparation des données

1. Sources des données intégrées

La construction de la base de données intégrée a nécessité l'exploitation de plusieurs sources biologiques afin de représenter de manière complète les gènes, leurs fonctions et leurs relations. Les informations sur les gènes de Escherichia coli K12 MG1655

comprennent leur identifiant unique, leur position chromosomique et le locus correspondant, permettant de positionner chaque élément dans le génome. Afin de faciliter la correspondance entre différentes bases de données, des alias ou synonymes des gènes ont également été intégrés, assurant ainsi une meilleure cohérence entre les identifiants issus de sources hétérogènes.

Les annotations fonctionnelles des protéines codées par ces gènes proviennent principalement de UniProt, via les keywords, qui permettent de caractériser les fonctions biologiques et les propriétés moléculaires des protéines. Ces informations ont été complétées par les domaines protéiques InterPro, fournissant des détails supplémentaires sur les structures et motifs fonctionnels. Pour relier les données aux connaissances scientifiques, des références bibliographiques PubMed ont été intégrées, permettant de contextualiser chaque gène ou protéine dans la littérature.

Les aspects régulateurs et métaboliques des gènes ont été pris en compte grâce aux unités transcriptionnelles et pathways d'EcoCyc, qui renseignent sur les ensembles de gènes coordonnés dans des voies métaboliques spécifiques. Les catégories fonctionnelles standardisées, représentées par les GO Terms (Gene Ontology), ont également été intégrées afin de normaliser l'annotation biologique des gènes et des protéines. Enfin, pour quantifier les relations entre gènes, des scores issus de StringDB ont été ajoutés, basés sur la co-expression, l'interaction protéine-protéine et les liens phylogénomiques, permettant d'évaluer les associations fonctionnelles entre paires de gènes ou protéines.

L'intégration de ces différentes sources a permis de créer une base de données riche, cohérente et exploitable, adaptée aux analyses d'enrichissement et à l'interrogation systématique des ensembles de gènes.

2. Librairies et programmes utilisés

- **Python (version 3.12.11)**

Utilisé pour interroger Neo4j, effectuer la recherche d'enrichissement et comparer les mesures statistiques.
Librairies utilisées : neo4j, argparse, os.path, json, pandas, scipy.stats, random, pathlib, time

- **R (version 4.4.1)**

Utilisé pour intégrer les données dans Neo4j, réaliser les statistiques descriptives et générer les figures d'enrichissement.
Librairies utilisées : neo2R, reticulate, tidyverse (incluant dplyr, readr, tidyr), ggplot2

- **Base de données Neo4j (neo4j:5.26.12-community)**

Utilisée pour structurer les gènes, annotations et relations sous forme de graphe et permettre des requêtes efficaces pour l'analyse d'enrichissement.

3. Choix et format des fichiers

Pour assurer une intégration cohérente et reproductible des différentes sources de données, nous avons choisi d'utiliser des fichiers au format TSV (Tab-Separated Values). Ce format simple et largement compatible permet de stocker les données sous forme de tableau, facilitant leur lecture, leur manipulation et leur import dans Neo4j via des scripts automatisés. Les colonnes ont été normalisées pour contenir des identifiants uniques, l'organisme concerné et la source de la donnée, garantissant ainsi une homogénéité entre les différents fichiers.

Chaque type de données a été traité selon ses spécificités. Les informations sur les gènes et leurs alias ont été extraites avec leurs identifiants principaux et secondaires afin de permettre une correspondance correcte avec les autres jeux de données. Les mots-clés UniProt, les domaines InterPro et les GO Terms ont été normalisés pour conserver uniquement les identifiants pertinents et standardisés, ce qui facilite leur utilisation dans les analyses d'enrichissement. Pour les références PubMed, seules les associations validées aux gènes ou protéines ont été conservées, de manière à éviter les doublons et les informations redondantes. Les pathways et unités transcriptionnelles ont été importés avec les listes de gènes correspondantes, permettant de représenter les ensembles de gènes comme des cibles dans la base. Enfin, les scores d'association StringDB ont été stockés sous forme de paires de gènes avec leurs valeurs de confiance, assurant une lecture directe par le moteur de graphes.

Cette approche garantit que chaque fichier d'entrée est exploitable, lisible et cohérent avec les autres sources, tout en facilitant la traçabilité des données originales. La normalisation des identifiants et des champs clés assure également la compatibilité avec les scripts Python développés pour l'analyse d'enrichissement et la comparaison des mesures de similarité.

4. Construction de la base Neo4j

La construction de la base de données Neo4j a consisté à intégrer l'ensemble des fichiers préparés afin de représenter les gènes et leurs relations sous forme de graphe. Chaque entité (gène, alias, mot-clé, domaine, pathway, TU, GO Term) est modélisée comme un nœud et les relations biologiques ou fonctionnelles entre ces entités sont représentées par des arêtes. Cette approche permet une interrogation flexible et rapide, notamment pour les analyses d'enrichissement.

4.1 Modèle de données (Schéma)

Le schéma de la base est illustré dans la figure suivante :

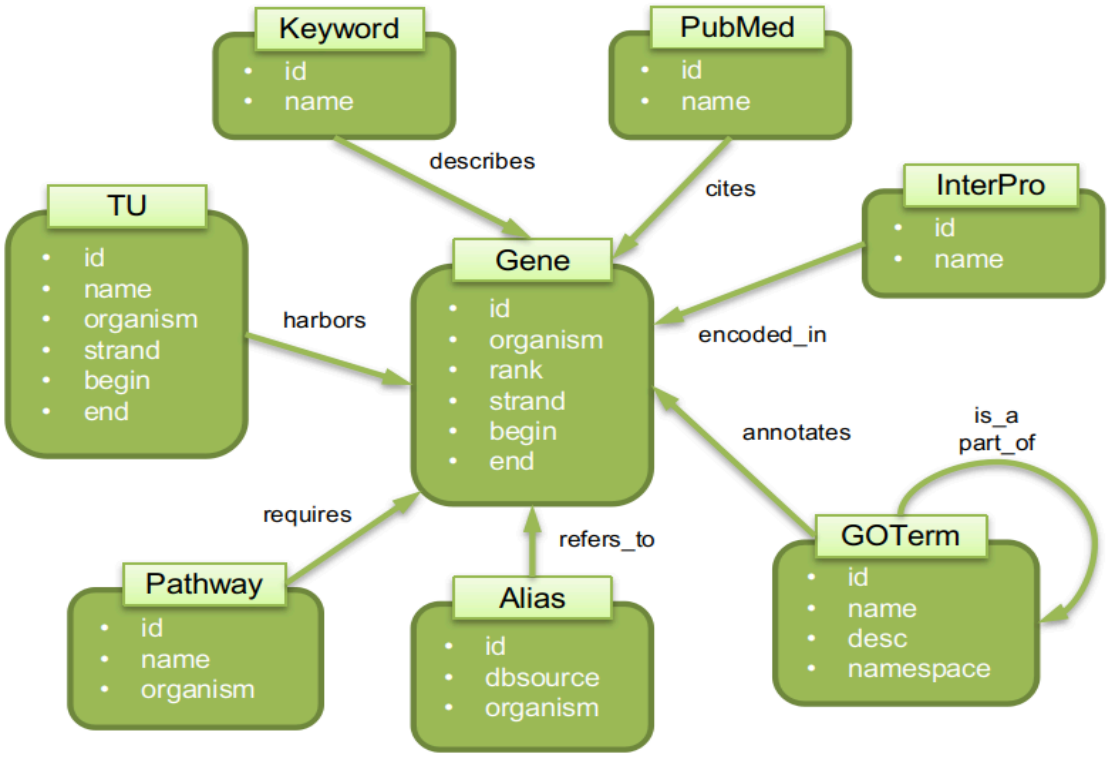

Figure 1: Schéma illustrant les sommets et liens de la base de données

Dans ce schéma, les gènes sont connectés à leurs alias et à leurs annotations fonctionnelles (GO Terms, InterPro, Keywords). Les pathways et unités transcriptionnelles relient plusieurs gènes afin de constituer des ensembles de référence pour l'analyse d'enrichissement. Les scores d'association (StringDB) sont représentés par des relations pondérées entre gènes, reflétant la force des liens biologiques ou fonctionnels.

Des contraintes et index ont été créés pour garantir l'intégrité et accélérer les requêtes. Les identifiants uniques des gènes et des mots-clés sont indexés afin de faciliter les recherches et d'éviter les doublons. Les relations ont été définies pour refléter le type biologique ou fonctionnel de l'association, en utilisant les types suivants : annotates, cites, describes, encoded_in, harbors, is_a, part_of, refers_to et requires. Chaque relation permet de représenter fidèlement le lien entre les différentes entités de la base (gènes, protéines, annotations fonctionnelles, pathways, unités transcriptionnelles, etc.).

4.2 Scripts d'import

L'import des données a été réalisé à l'aide de commandes Cypher. Chaque fichier TSV a été chargé en créant d'abord les nœuds correspondants, puis en établissant les relations entre ces nœuds. Des précautions ont été prises pour gérer les doublons éventuels et les alias multiples, en utilisant les identifiants normalisés comme clé de correspondance.

Certaines difficultés ont été rencontrées lors de l'import, principalement liées à la gestion des alias. Plusieurs synonymes apparaissent associés à différents bnumbers selon les sources ; plutôt que de les supprimer, ils ont été conservés et considérés comme des alias valides pour tous les gènes concernés, afin de préserver l'intégralité de l'information. En revanche, les alias contenant déjà un identifiant de type bxxxx ont été filtrés, car ils correspondaient à un identifiant canonique et auraient créé des relations redondantes et non informatives entre un gène et un alias identique à son propre identifiant. Ce prétraitement a permis d'éviter les duplications inutiles tout en maintenant une représentation fidèle des correspondances entre identifiants. Les scripts d'import ont été structurés pour gérer systématiquement ces cas et garantir une reconstruction cohérente, complète et reproductible de la base Neo4j.

5. Statistiques descriptives de la base intégrée

L'ensemble des fichiers TSV préparés a été importé dans la base Neo4j, produisant un graphe riche couvrant les gènes, leurs annotations fonctionnelles et leurs relations biologiques. Une première exploration descriptive a été réalisée afin d'évaluer la structure générale de la base et la densité des connaissances intégrées.

La base contient 4318 gènes de Escherichia coli K-12 MG1655, auxquels sont associés 12 249 alias, ce qui reflète la forte hétérogénéité des synonymes présents dans les bases d'annotation. Les aspects fonctionnels sont particulièrement bien couverts : la base compte 39 906 termes GO, 384 mots-clés UniProt, 7 736 domaines InterPro, ainsi que 435 pathways et 2572 unités transcriptionnelles (TU). La dimension bibliographique est également riche, avec 35 320 références PubMed intégrées.

Les relations entre entités sont nombreuses et variées. On dénombre 22 587 associations Gene-GO, complétées par les relations hiérarchiques internes de la Gene Ontology : 61 712 relations "is_a" et 6 598 relations "part_of" entre termes GO. Les liens fonctionnels de type Keyword-Gene représentent 27 621 relations, tandis que la base contient 11 916 liens Alias-Gene, reflétant les synonymes multiples des gènes. Les aspects biologiques structuraux sont couverts à travers 2 723 relations Pathway-Gene, 7 818 relations TU-Gene, et 17 176 associations InterPro-Gene. Les références bibliographiques sont largement connectées aux gènes avec 101 841 relations PubMed-Gene. Enfin, les interactions issues de StringDB fournissent un réseau dense d'associations entre gènes, totalisant 496 856 arêtes Gene-Gene, ce qui constitue la structure relationnelle la plus riche de la base.

Ces statistiques confirment que la base intégrée couvre à la fois les dimensions fonctionnelle, structurale, interactionnelle et bibliographique des connaissances disponibles sur E. coli, fournissant ainsi un support solide pour l'analyse et la comparaison des ensembles de gènes réalisée dans la suite du projet.

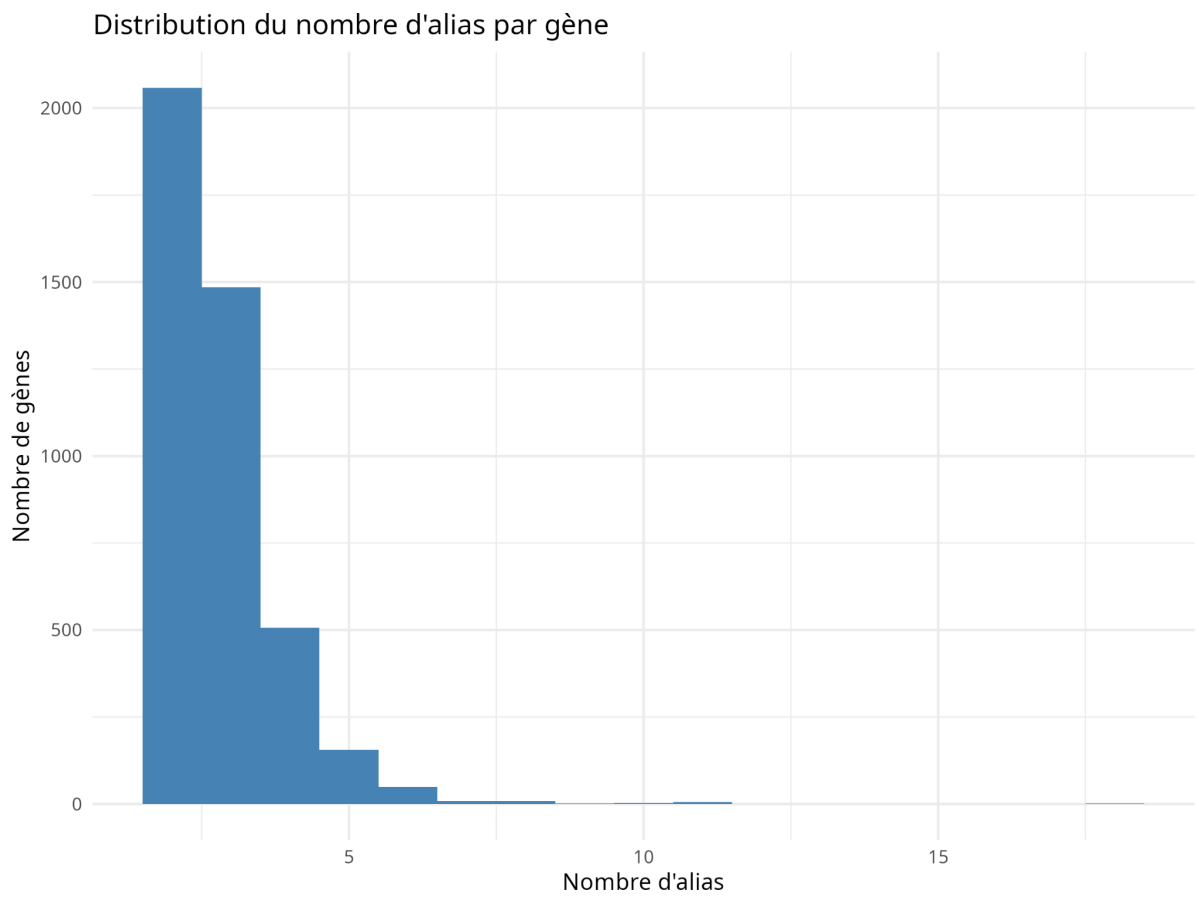

Figure 2 : Distribution du nombre d’alias par gène

La distribution montre que la majorité des gènes possèdent un faible nombre d’alias. Plus de 2000 gènes disposent de 2 alias, ce qui constitue le mode de la distribution. Environ 1500 gènes possèdent 3 alias, tandis que les gènes ayant 4 ou 5 alias sont beaucoup moins nombreux. Les valeurs supérieures sont quasi absentes, ce qui suggère que les alias sont relativement bien normalisés et concentrés autour de quelques termes alternatifs par gène.

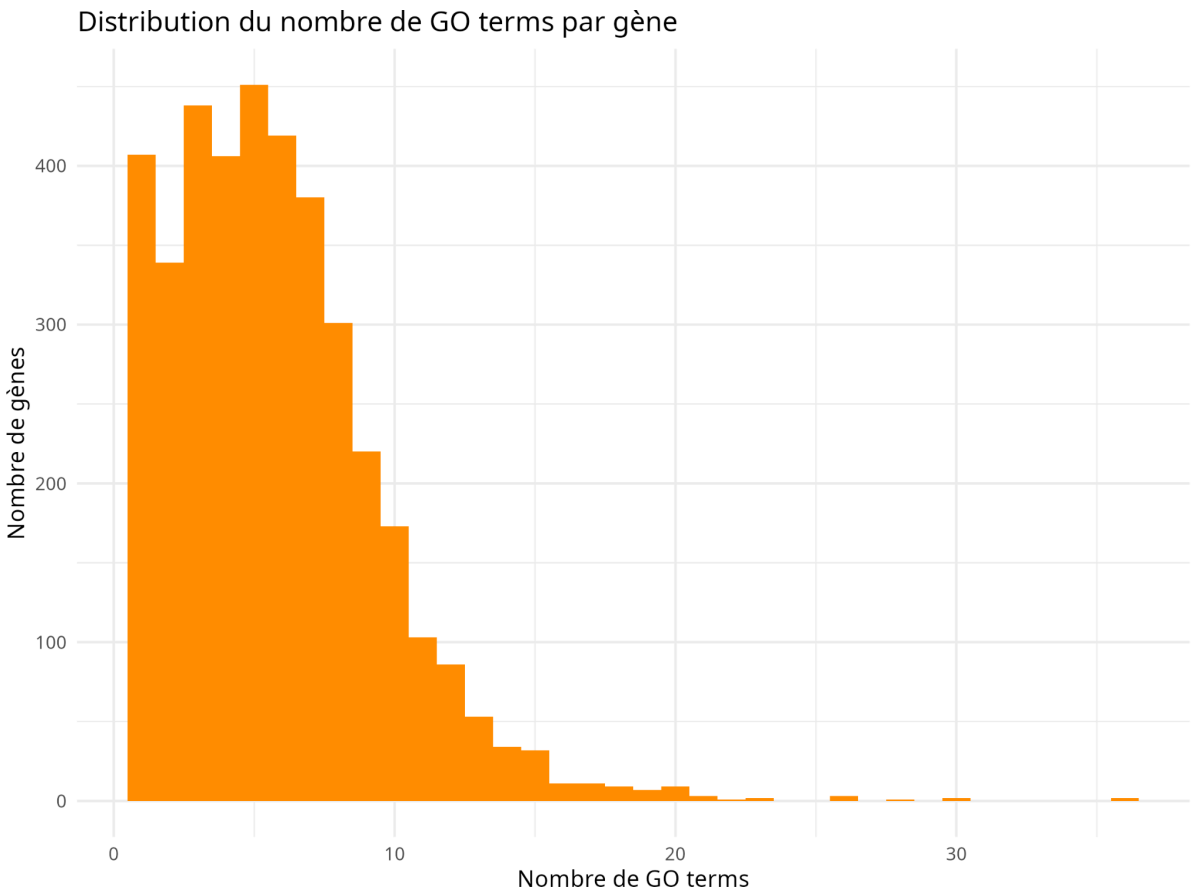

Figure 3 : Distribution du nombre de termes GO par gène

La majorité des gènes sont annotés par un nombre restreint de termes GO, généralement compris entre 1 et 10. Au-delà de 10 GO terms, la fréquence décroît fortement, indiquant que seuls quelques gènes, probablement multifonctionnels ou très étudiés, possèdent un grand nombre d'annotations fonctionnelles.

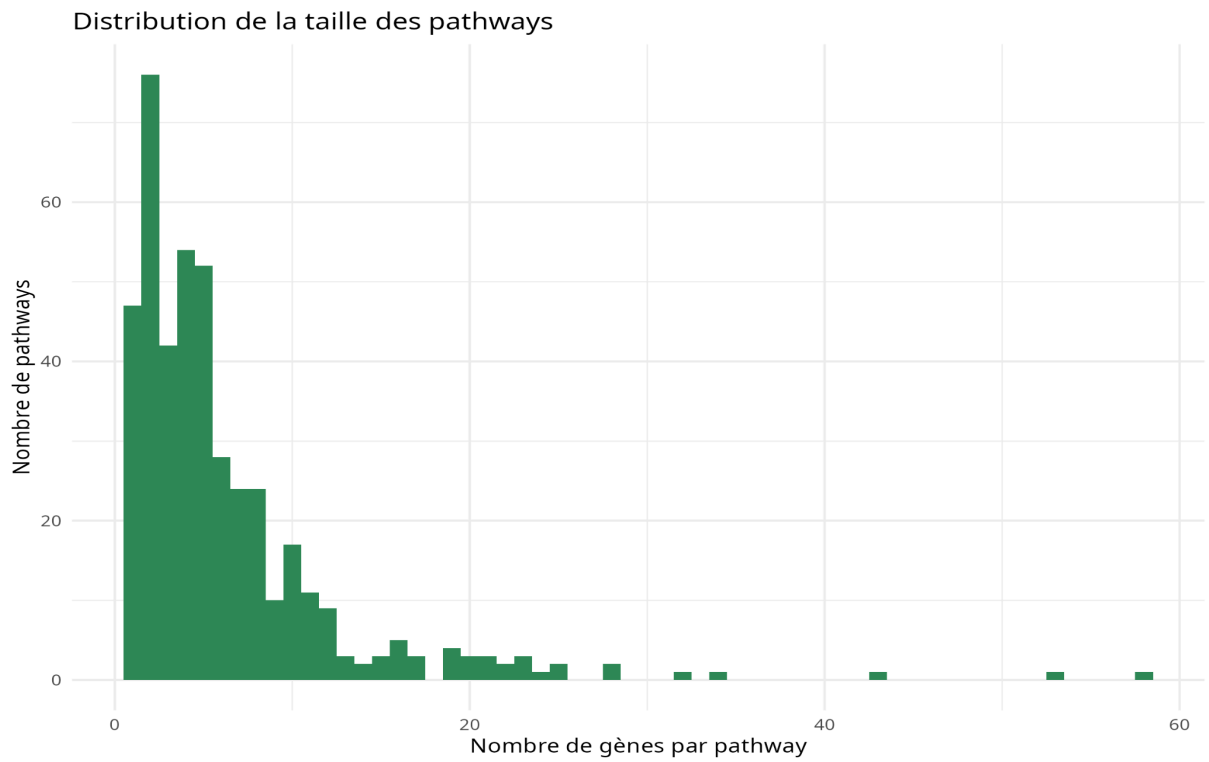

Figure 4 : Taille des pathways (nombre de gènes par pathway)

Les pathways présentent majoritairement une taille réduite : la majorité d'entre eux mobilisent moins de 10 gènes. Cette distribution suggère une organisation fonctionnelle en modules spécialisés et relativement compacts plutôt qu'en larges réseaux métaboliques.

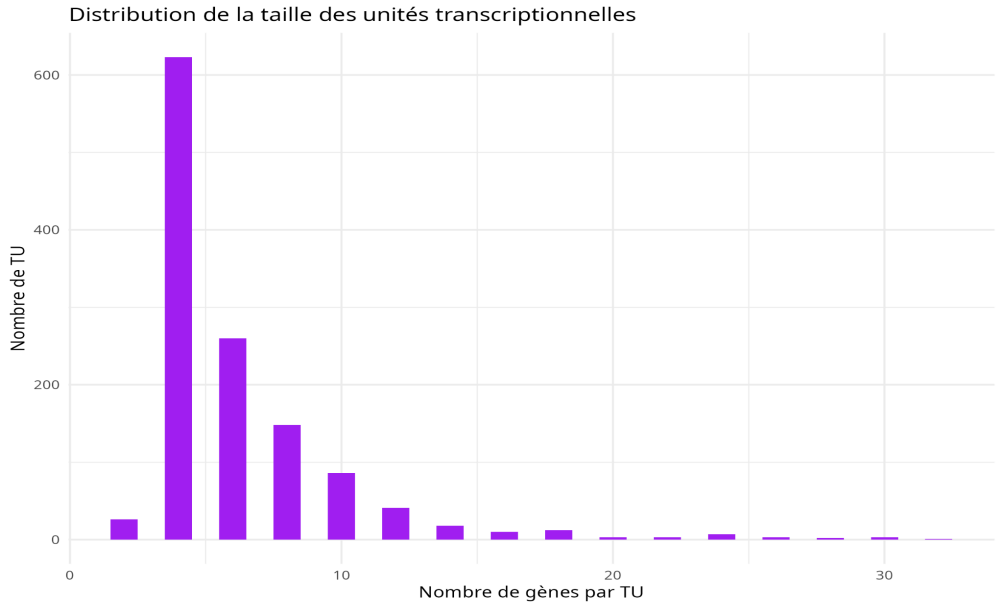

Figure 5 : Taille des unités transcriptionnelles (TU)

La plupart des unités transcriptionnelles contiennent entre 4 et 12 gènes. Cela montre une organisation transcriptionnelle structurée autour de petits groupes de gènes co-transcrits, ce qui est cohérent avec la logique opéronique caractéristique des bactéries comme E. coli.

III. Ajout de fonctionnalités au script d'enrichissement

L'objectif initial du script Python fourni était de réaliser une analyse d'enrichissement sur un ensemble de gènes, en identifiant les catégories fonctionnelles surreprésentées dans la liste fournie. Cependant, la version d'origine présentait plusieurs limitations : elle ne proposait qu'une seule mesure statistique (la loi binomiale), n'offrait pas d'ajustement pour tests multiples, ne retournait pas d'informations descriptives sur les ensembles enrichis et ne permettait pas une exploitation structurée des résultats. Afin de répondre aux besoins du projet et d'améliorer la robustesse de l'analyse, plusieurs fonctionnalités ont été ajoutées au script.

Tout d'abord, le script a été étendu pour intégrer quatre mesures statistiques différentes, en plus de la loi binomiale déjà disponible. La nouvelle version permet désormais d'utiliser la loi hypergéométrique, largement utilisée en analyse fonctionnelle, ainsi qu'un test d'indépendance χ^2, permettant de vérifier si l'appartenance des gènes au query est indépendante de leur association à une fonction donnée. Une métrique supplémentaire dite de couverture a également été introduite ; elle permet de quantifier le degré d'intersection relative entre l'ensemble fourni et chaque ensemble fonctionnel. Ces nouvelles mesures élargissent les possibilités d'analyse et permettent de sélectionner la méthodologie la plus adaptée à la nature des données étudiées.

Ensuite, la récupération des ensembles fonctionnels annotés a été améliorée. Alors que le script initial ne renvoyait que les identifiants associés, la version modifiée extrait désormais à la fois l'identifiant, la description de l'annotation et la liste des gènes correspondants. Cela permet d'obtenir un tableau de résultats beaucoup plus complet et exploitable, facilitant l'interprétation biologique des enrichissements détectés. Par ailleurs, les résultats sont désormais rassemblés dans une structure pandas DataFrame, ce qui permet un tri simple par p-value et une mise en forme plus lisible lors de l'affichage final.

Une autre amélioration importante concerne la gestion des p-values. Le script propose à présent une option permettant d'appliquer une correction des tests multiples selon la méthode de Benjamini-Hochberg (FDR). Cette fonctionnalité, absente de la version d'origine, est essentielle pour éviter les faux positifs lorsque de nombreux ensembles sont testés simultanément. Une option supplémentaire permet également de limiter le nombre de résultats affichés, ce qui facilite la lecture lorsque seules les annotations les plus significatives sont d'intérêt.

Enfin, l'ensemble de la gestion des requêtes Cypher et de la connexion à Neo4j a été amélioré. Le script utilise désormais le transformateur neo4j.Result.to_df, ce qui renvoie directement les résultats sous forme de DataFrame. La connexion est explicitement vérifiée puis correctement fermée à la fin de l'exécution, ce qui renforce la stabilité et la reproductibilité de l'analyse.

L'utilisation du script modifié reste simple. L'utilisateur fournit une liste de gènes (sous forme de fichier ou directement en argument) ainsi que le type d'annotations à tester (Keywords, GOTerms, Pathways…). Plusieurs options permettent de personnaliser l'analyse, notamment le choix de la mesure statistique (--measure), l'application d'une correction FDR (--adjust), l'ajustement du seuil de significativité (--alpha) et la limitation du nombre de sorties (--limit). Par exemple, la commande suivante permet de rechercher des GO terms enrichis en utilisant la loi hypergéométrique et une correction FDR : **python3 blastsets.neo4j.py -q genes.txt -t GOTerm -m hypergeometric -c -v**
De même, un enrichissement sur les Keywords avec la fonction coverage peut être réalisé via : **python3 blastsets.neo4j.py -q genes.txt -t Keyword -m coverage -v**

En résumé, ces modifications apportent une amélioration significative du script original, tant en termes de fonctionnalités statistiques que de qualité du rendu. Elles permettent désormais de réaliser une analyse d'enrichissement complète, fiable et adaptée aux besoins du projet.

IV. Comparaison des mesures intégrées

1. Approches envisagées et méthode pour en sélectionner une

La comparaison des mesures intégrées avait pour objectif d'identifier celles offrant les performances les plus fiables pour retrouver la bonne catégorie fonctionnelle à partir d'un ensemble de gènes enrichi. Plusieurs approches étaient envisageables, mais l'évaluation a reposé sur une stratégie empirique. Pour cela, différentes configurations de simulations ont été générées en faisant varier la taille des ensembles de gènes, la proportion de signal réel et le niveau de bruit ajouté. À chaque simulation, les quatre mesures binomiale, hypergéométrique, chi² et coverage ont été appliquées indépendamment, puis leurs résultats ont été comparés à la catégorie attendue.

La mesure la plus performante devait être celle qui parvenait le plus fréquemment à classer la bonne catégorie dans les premières positions, tout en conservant une stabilité de comportement sur l'ensemble des scénarios testés. L'évaluation s'est donc appuyée sur des indicateurs tels que la proportion de cas où la bonne catégorie apparaît en première position, sa fréquence dans les cinq meilleurs résultats et la qualité du classement global (rank médian, MRR). Cette approche systématique permet de sélectionner une mesure non pas sur des critères théoriques, mais sur la base d'une observation directe de son comportement.

2. Mise en œuvre

La mise en œuvre a consisté à appliquer les quatre mesures sur plusieurs dizaines de scénarios simulés. Chaque simulation associait un ensemble de gènes à une catégorie fonctionnelle cible, puis injectait un niveau variable de bruit et de signal. Les quatre méthodes étaient ensuite appliquées de manière identique, sans ajustement spécifique, afin de garantir la comparabilité des résultats.

Pour chaque mesure, les positions obtenues par la catégorie cible étaient enregistrées. L'ensemble des simulations a ainsi permis d'agréger des statistiques globales révélant non seulement les performances moyennes des différentes approches, mais aussi leur stabilité et leur sensibilité aux variations de taille ou de bruit. Cette procédure permet une comparaison fine et robuste entre les mesures.

3. Synthèse des résultats obtenus

Les résultats mettent en évidence une hiérarchie claire entre les quatre mesures évaluées. Les mesures probabilistes basées sur des modèles de comptage : binomiale et hypergéométrique se distinguent nettement par leurs performances globales. Elles montrent toutes deux une forte capacité à identifier correctement la catégorie cible, avec une proportion élevée de classements en première ou en très bonne position, ainsi qu'un rang médian particulièrement faible. Ces deux approches se révèlent également très stables, quels que soient la taille de la requête ou le niveau de bruit.

La mesure chi² présente des résultats intermédiaires : elle reste performante dans les cas où les ensembles impliqués sont de grande taille, mais ses performances se dégradent dès que le signal devient plus faible ou que les tailles sont déséquilibrées. Sa variabilité plus élevée la rend globalement moins fiable que les deux mesures précédentes.

Enfin, la mesure de coverage s'avère la moins performante dans presque toutes les conditions. Bien qu'elle soit intuitive et facile à interpréter, elle se montre très sensible au bruit et aux différences de taille entre catégories, ce qui entraîne des classements moins précis et nettement moins stables.

Les tendances décrites ci-dessus sont illustrées par la figure ci-dessous, qui compare les performances des quatre mesures d'enrichissement selon la taille des requêtes et l'intensité du signal. Cette représentation visuelle met en évidence les écarts observés au cours des expérimentations, notamment dans les conditions de faible signal où les différences entre

approches sont les plus marquées.

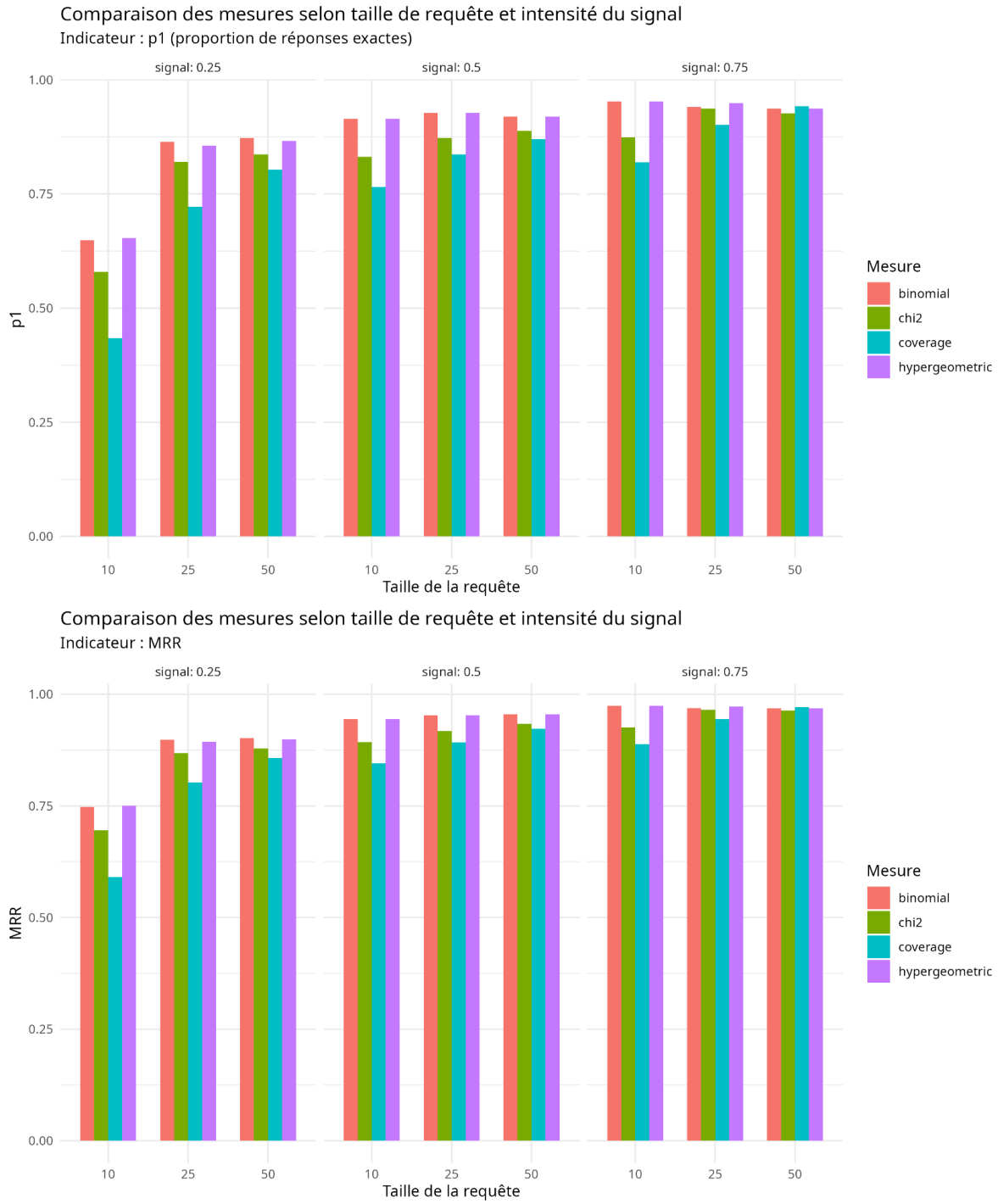

Figure 6 : Comparaison des performances des quatre mesures d'enrichissement selon la taille de la requête et l'intensité du signal (indicateurs p1 et MRR)

4. Bilan : quelle mesure utiliser de préférence et pourquoi ?

Au regard de l'ensemble des résultats, la mesure qui se démarque le plus clairement est l'hypergéométrique. Elle combine à la fois d'excellentes performances, une grande robustesse et un comportement cohérent avec les méthodes classiques d'enrichissement utilisées en biologie. Sa capacité à maintenir un bon classement de la catégorie cible même dans des scénarios difficiles en fait une approche fiable et adaptée.

La mesure binomiale constitue une alternative tout aussi solide, avec des performances très proches, parfois équivalentes. Elle peut être privilégiée lorsque la simplicité de calcul est un critère important ou comme méthode complémentaire offrant une validation des résultats obtenus par l'hypergéométrique.

En revanche, les mesures chi² et coverage apparaissent moins adaptées : la première souffre d'une variabilité notable, la seconde d'une sensibilité excessive au bruit et à la structure des catégories.

5. Perspectives d'amélioration

Plusieurs pistes d'amélioration émergent de cette comparaison. Un premier axe concerne la possibilité d'affiner la modélisation statistique, par exemple en intégrant des variantes de l'hypergéométrique ou des tests exacts mieux adaptés aux petites tailles. Une seconde perspective consiste à équilibrer davantage les catégories fonctionnelles ou à introduire des simulations dont le bruit reflète mieux des situations biologiques réelles, afin de tester la robustesse des mesures dans des contextes plus complexes.

Il serait également pertinent d'explorer des approches hybrides combinant plusieurs mesures, notamment l'hypergéométrique et la binomiale, afin de produire des scores plus stables ou un consensus entre méthodes. Enfin, l'intégration de techniques de correction ou de pondération pourrait réduire la sensibilité observée pour certaines mesures, en particulier le chi² et le coverage, et améliorer leur comportement global.

V. Analyse de l'ensemble de gènes fourni

L'ensemble de gènes set.M2.14.txt et set.02.txt ont été examinés afin de caractériser les fonctions biologiques prédominantes, les motifs protéiques enrichis, les mots-clés associés et les unités transcriptionnelles (TU). Chaque membre du binôme a réalisé l'analyse de son jeux de données : l'un a effectué les enrichissements GO, Keyword, InterPro, Pathway et TU pour l'ensemble de gènes set.M2.14.txt , tandis que l'autre s'est concentré sur GO, Keyword, InterPro, Pathway et PubMed pour l'ensemble de gènes set.02.txt. Cette approche permet de comparer les interprétations et de couvrir à la fois la fonction générale et les modules spécialisés de ces ensembles. L'objectif est de mettre en évidence les processus biologiques dominants et les relations potentielles entre gènes.

A. Ensemble de gènes d'intérêt set.M2.14.txt

1. Résultats obtenus par la recherche d'enrichissement

1.1 GOTerm

L'analyse des termes GO montre que les gènes de l'ensemble set.M2.14.txt sont fortement enrichis pour des fonctions liées à la traduction. Les termes les plus significatifs sont cytoplasmic translation, structural constituent of ribosome, rRNA binding et translation, avec des p-values extrêmement faibles, indiquant une sur-représentation très nette. Cette observation confirme que la majorité des gènes codent pour des protéines ribosomales et des facteurs de traduction. Des processus secondaires, tels que l'assemblage des sous-unités ribosomales, sont également représentés, suggérant une coordination fine des étapes de la traduction.

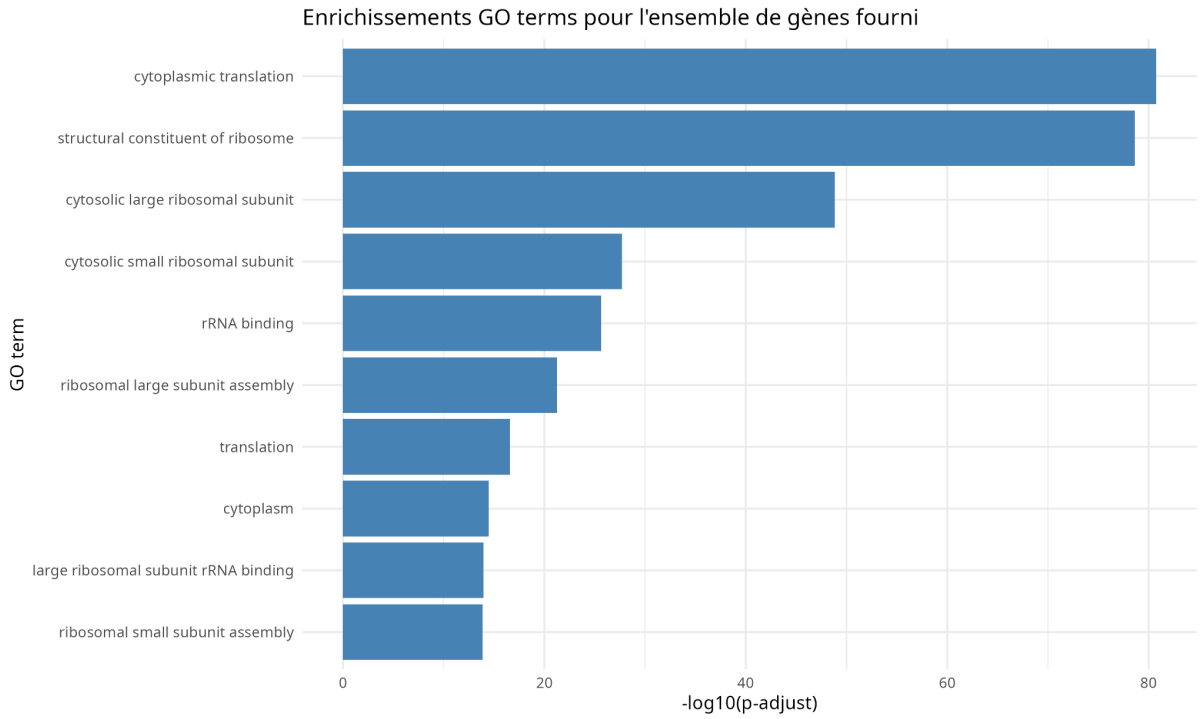

Figure 7 : Enrichissements des termes GO pour l'ensemble de gènes fourni

1.2 InterPro

L'enrichissement des domaines protéiques met en évidence plusieurs motifs fonctionnels clés. La voie de translocation Twin-arginine est particulièrement représentée, ainsi que les domaines molybdopterin oxidoreductase et les sous-unités [NiFe]-hydrogenase, qui interviennent dans le métabolisme énergétique et la respiration anaérobie. Les protéines ribosomales apparaissent également comme significativement enrichies, renforçant le rôle central de la traduction. Ces résultats soulignent la combinaison d'activités traductionnelles

et métaboliques spécialisées, avec un accent sur les fonctions redox et la respiration anaérobie.

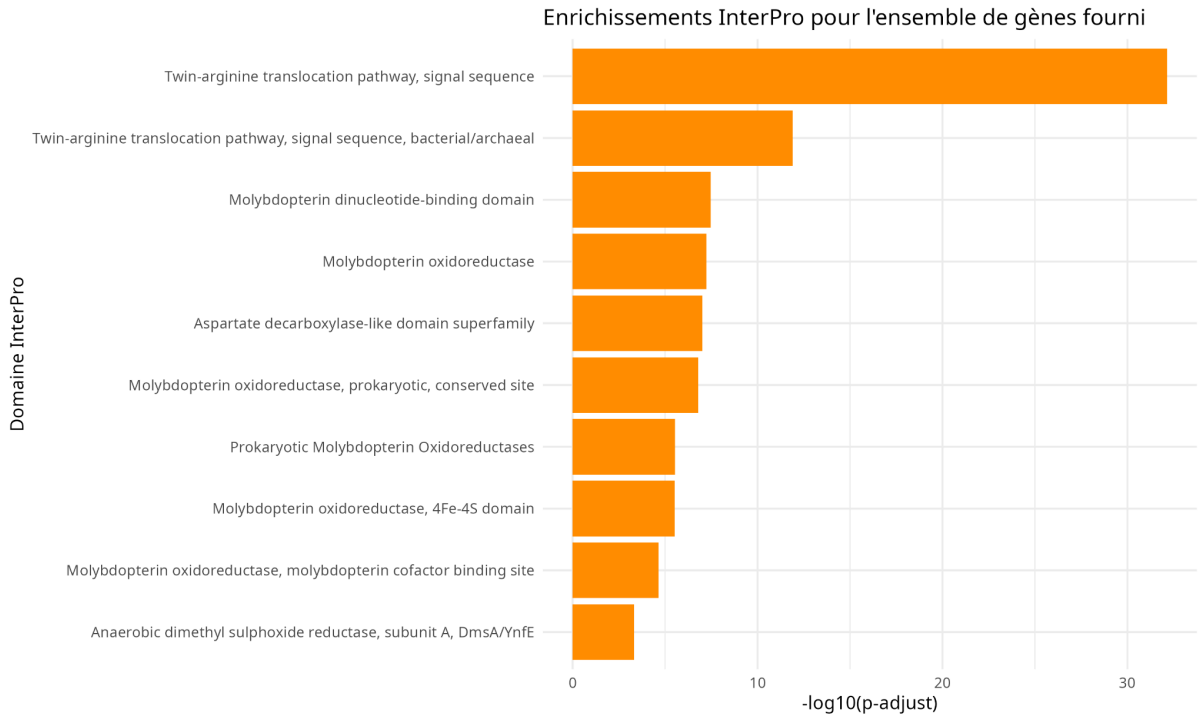

Figure 8 : Enrichissements des domaines protéiques InterPro

1.3 Keyword

Les mots-clés enrichis reflètent la diversité fonctionnelle de l'ensemble. On observe une sur-représentation de Ribosomal protein, RNA-binding, tRNA-binding, confirmant l'importance des fonctions de traduction. Parallèlement, des mots-clés comme Molybdenum, 4Fe-4S, Oxidoreductase identifient des gènes impliqués dans le métabolisme énergétique et la respiration anaérobie. Cette combinaison suggère que l'ensemble de gènes n'est pas uniquement centré sur la traduction, mais contient également des gènes participant à des voies métaboliques spécialisées.

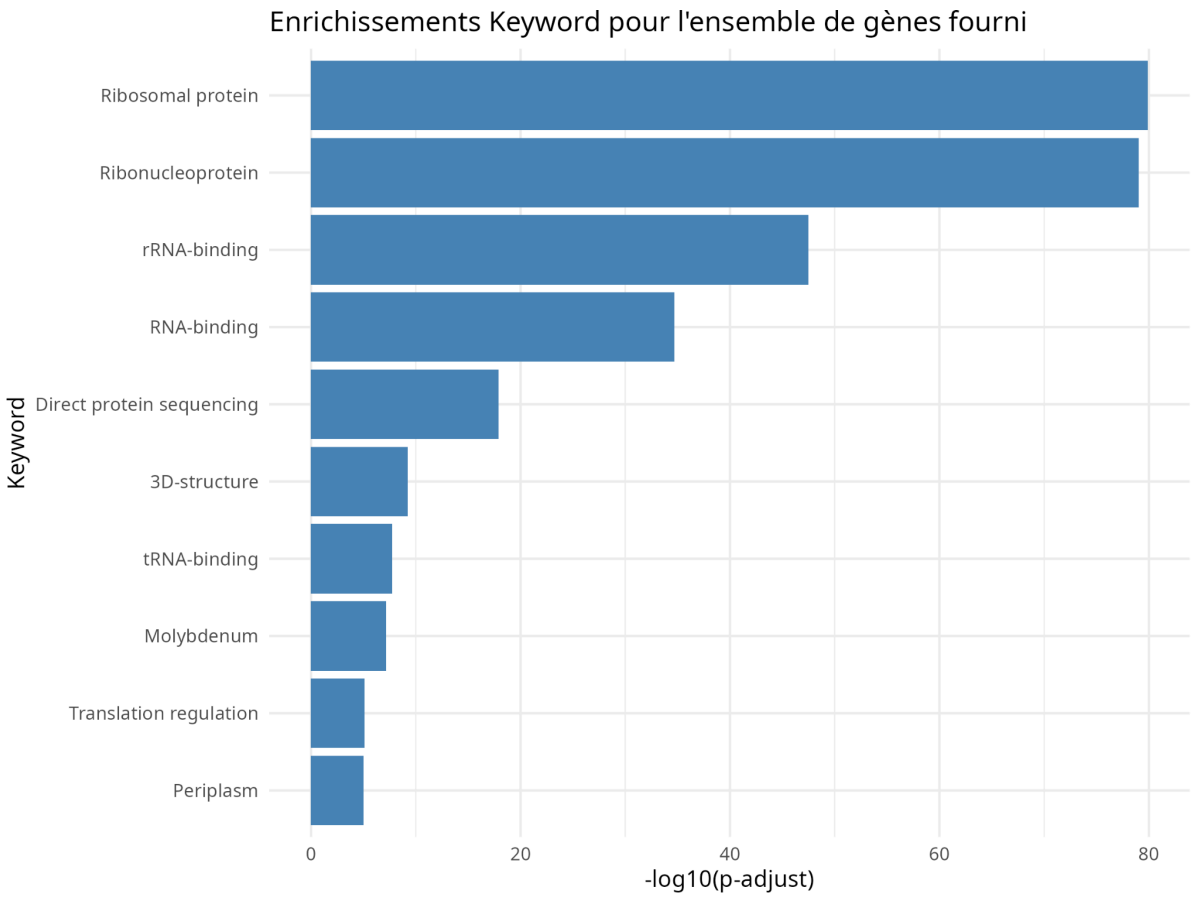

Figure 9 : Enrichissements des mots-clés fonctionnels

1.4 Pathway

L'analyse des voies métaboliques enrichies montre des transferts d'électrons spécifiques. Les voies principales incluent :

- Hydrogène vers triméthylamine N-oxyde ou diméthylsulfoxyde,
- Formate vers ces accepteurs ou vers le nitrite,
- Respiration anaérobie du sn-glycérol 3-phosphate.

Ces résultats confirment la participation de certains gènes à des fonctions métaboliques spécialisées et à la respiration anaérobie, ce qui complète les observations des analyses InterPro et Keyword.

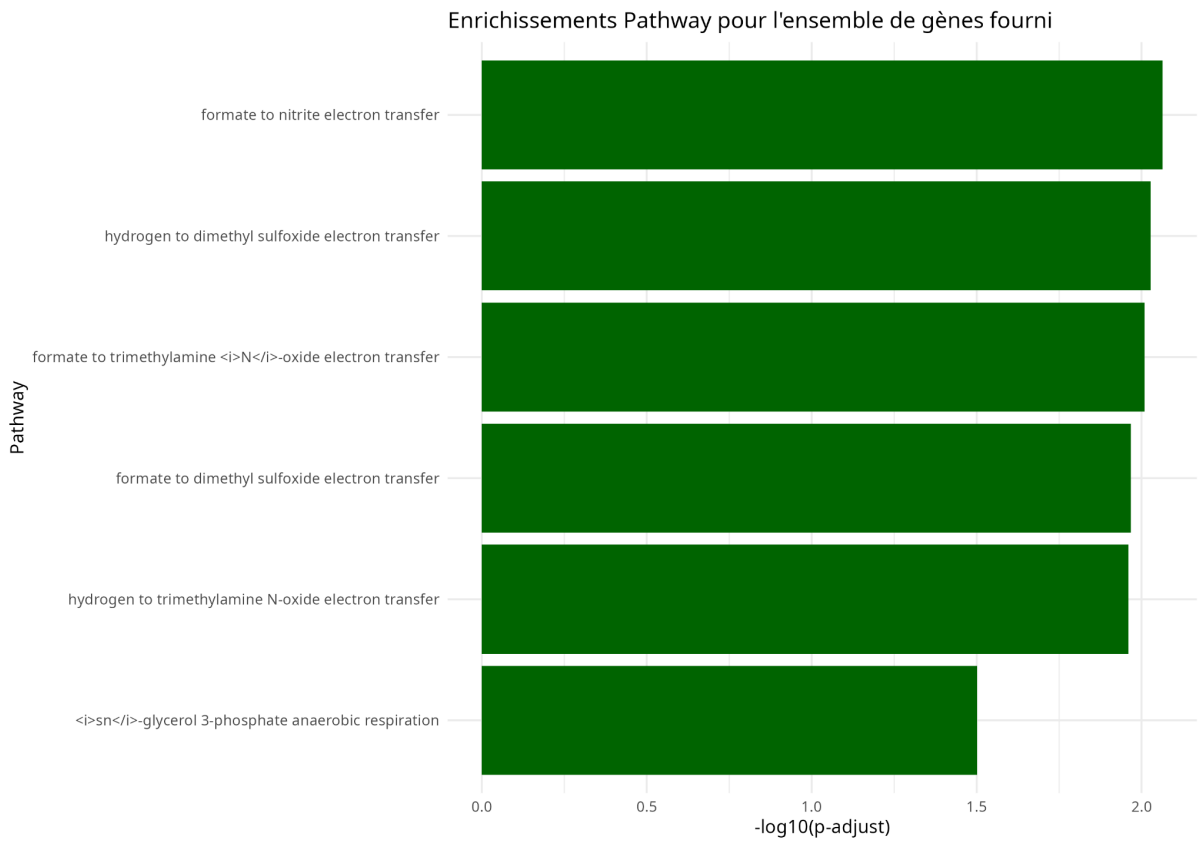

Figure 10 : Enrichissements des voies métaboliques (Pathway)

1.5 Transcription Units (TU)

L'analyse des unités transcriptionnelles montre une organisation cohérente des gènes fonctionnellement associés. Plusieurs TU, comme TU00337 et TU00347, regroupent essentiellement des gènes ribosomaux, tandis que d'autres contiennent des gènes liés à la respiration anaérobie et au transport Tat. Cette organisation suggère une régulation transcriptionnelle coordonnée, optimisant l'expression simultanée de gènes participant à des fonctions biologiques connexes.

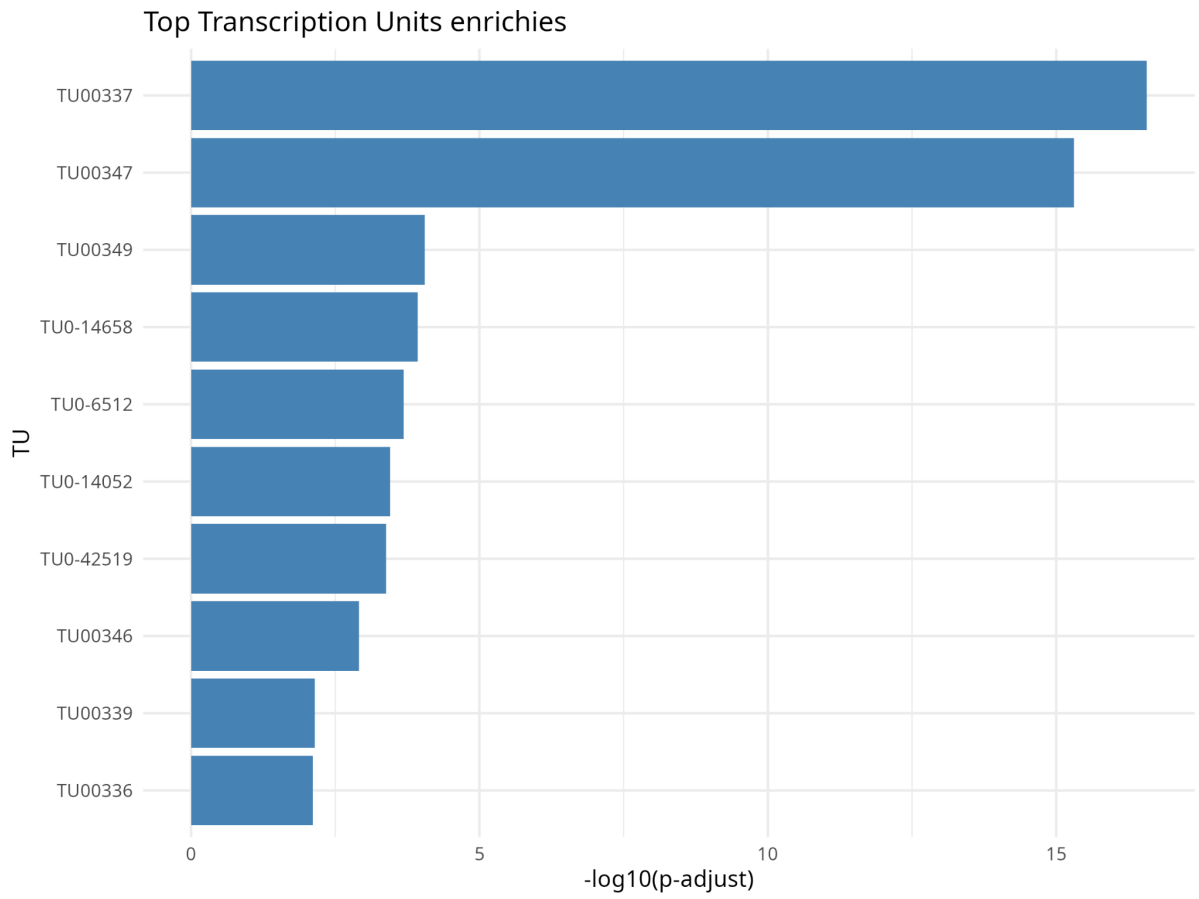

Figure 11 : Enrichissements des unités de transcription (TU)

2. Discussion et commentaires

L'ensemble set.M2.14.txt se distingue par un profil fonctionnel centré sur la traduction et les ribosomes, avec un sous-ensemble spécialisé dans la respiration anaérobie et le transfert d'électrons. Les enrichissements GOTerm et Keyword confirment la prépondérance des protéines ribosomales et de la machinerie de traduction, tandis que les analyses InterPro et Pathway identifient des motifs et des voies impliqués dans le métabolisme énergétique spécifique. L'organisation des gènes en TU renforce l'idée d'une cohérence fonctionnelle et transcriptionnelle au sein du génome.

Les figures fournissent une illustration claire de la répartition fonctionnelle : les enrichissements GO et Keyword mettent en évidence la traduction, InterPro et Pathway montrent les modules métaboliques spécialisés, et les TU illustrent la co-localisation et la régulation coordonnée des gènes. L'ensemble représente donc un exemple typique de regroupement fonctionnel de gènes ribosomaux avec des modules métaboliques spécialisés, pertinent pour des études sur la traduction et le métabolisme anaérobie.

B. Ensemble de gènes d'intérêt set.02.txt

1. Résultats obtenus par la recherche d’enrichissement

1.1 GOTerm

Les trois GO terms les plus enrichis dans le set set.02.txt sont cell division, regulation of cell shape, et peptidoglycan biosynthetic process, avec des p-values ajustées très faibles, indiquant un enrichissement significatif. Ces résultats montrent que les gènes du set participent principalement aux mécanismes de division cellulaire et à la construction de la paroi bactérienne, notamment via la synthèse du peptidoglycane. Les autres termes enrichis renforcent cette tendance et confirment que l’ensemble est fortement associé à la formation du septum et au maintien de la morphologie cellulaire.

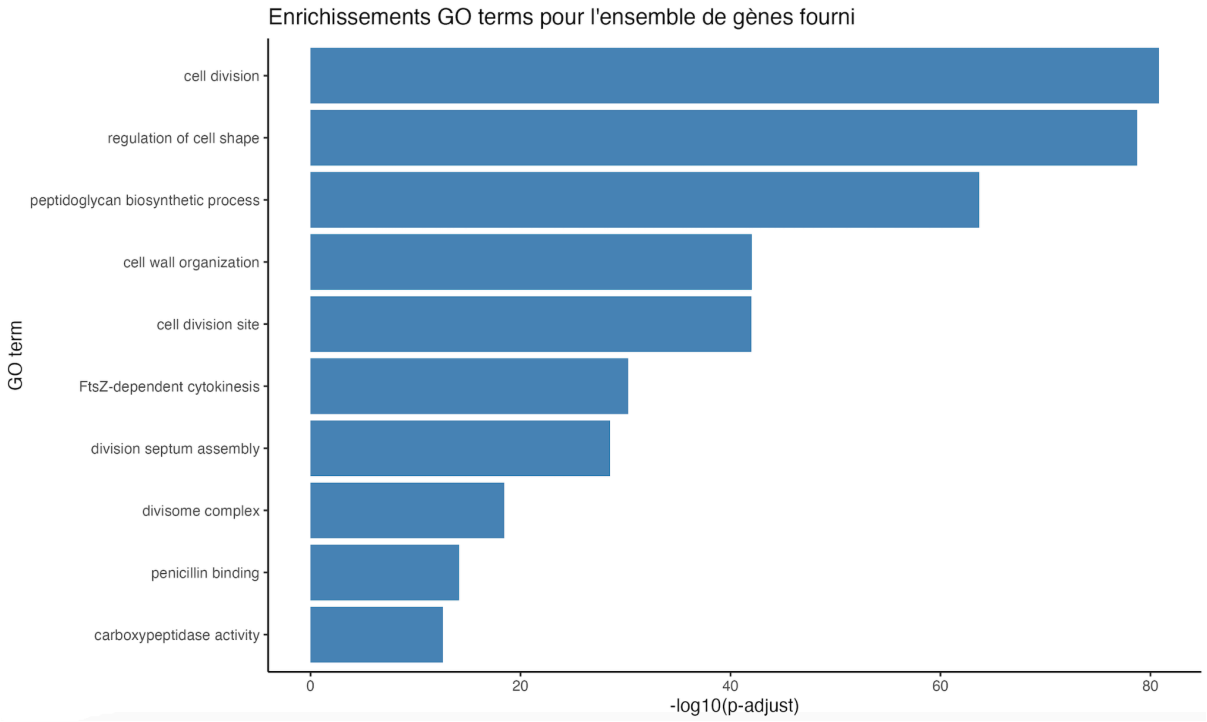

Figure 12: Enrichissements GO Terms de set.02.txt

1.2 InterPro

Les domaines InterPro les plus enrichis dans le set appartiennent majoritairement aux familles transpeptidases, L,D-transpeptidases et Mur ligases. Ces protéines sont caractéristiques de la biosynthèse du peptidoglycane, en particulier des étapes d’assemblage et de réticulation de la paroi. La présence de domaines liés aux penicillin-binding proteins (PBPs) confirme également l’implication de gènes ciblés par les β-lactamines et essentiels à la formation du septum.

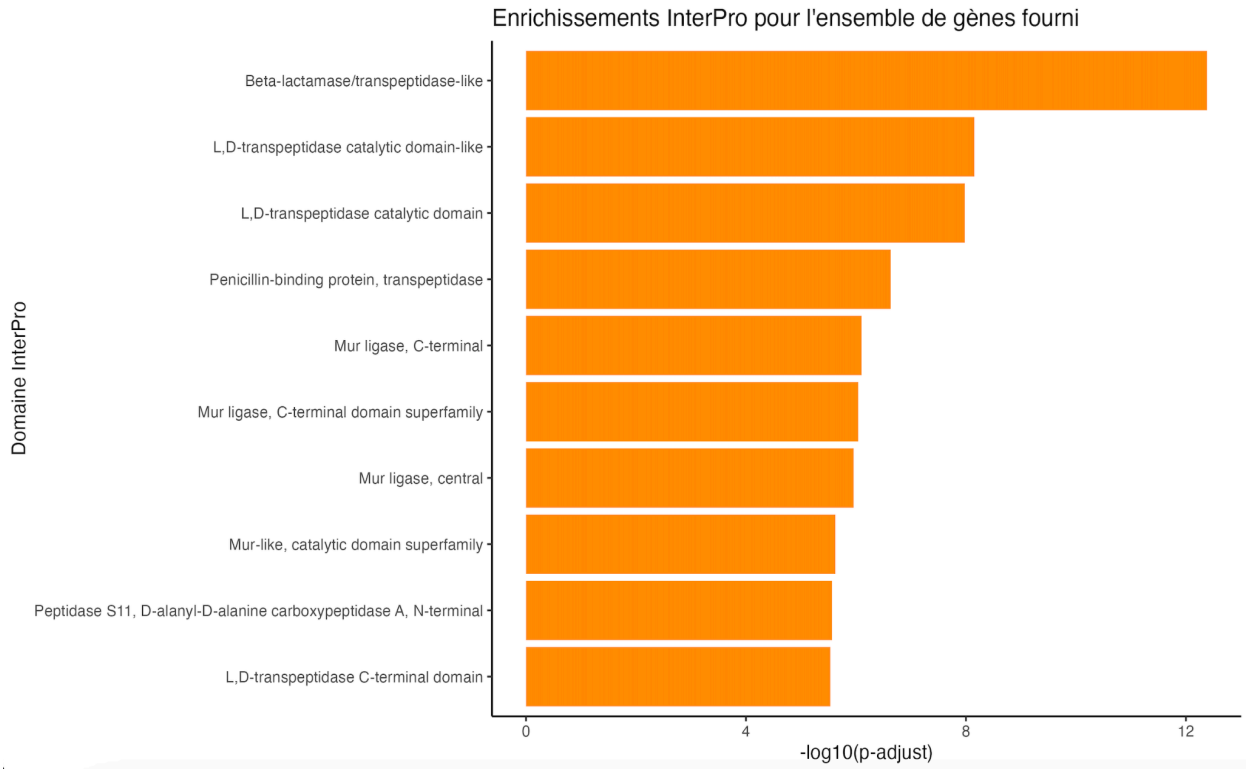

Figure 13: Enrichissements InterPro de set.02.txt

1.3 Keyword

Les keywords les plus enrichis dans set.02.txt confirment la forte implication des gènes dans la division cellulaire (Cell cycle, Cell division) et le maintien de la morphologie bactérienne (Cell shape). Les termes liés à la synthèse du peptidoglycane (Peptidoglycan synthesis, Cell wall biogenesis/degradation) renforcent l'idée que de nombreuses protéines du set participent à la construction et au remodelage de la paroi.

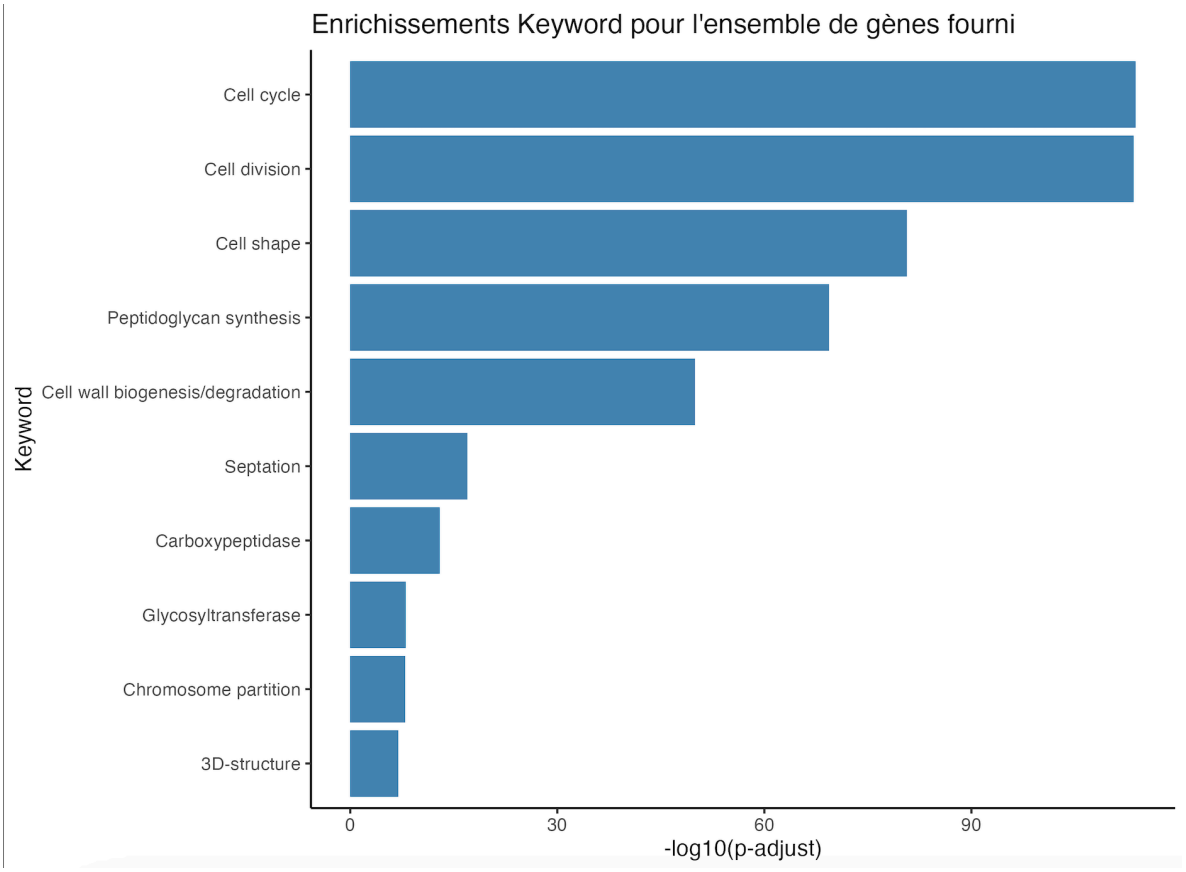

Figure 14: Enrichissements Keyword de set.02.txt

1.4 Pathway

Les pathways les plus enrichis dans set.02.txt concernent exclusivement la biosynthèse et la maturation du peptidoglycane, en particulier la voie meso-diaminopimelate containing, typique des bactéries à paroi Gram-négative. Ces voies regroupent les étapes clés de formation du précurseur du peptidoglycane, son assemblage et son recyclage. Les pathways liés à l'undécaprényl phosphate et aux polyisoprénoïdes renforcent l'idée que plusieurs gènes du set participent au transport et à l'ancrage des unités de la paroi.

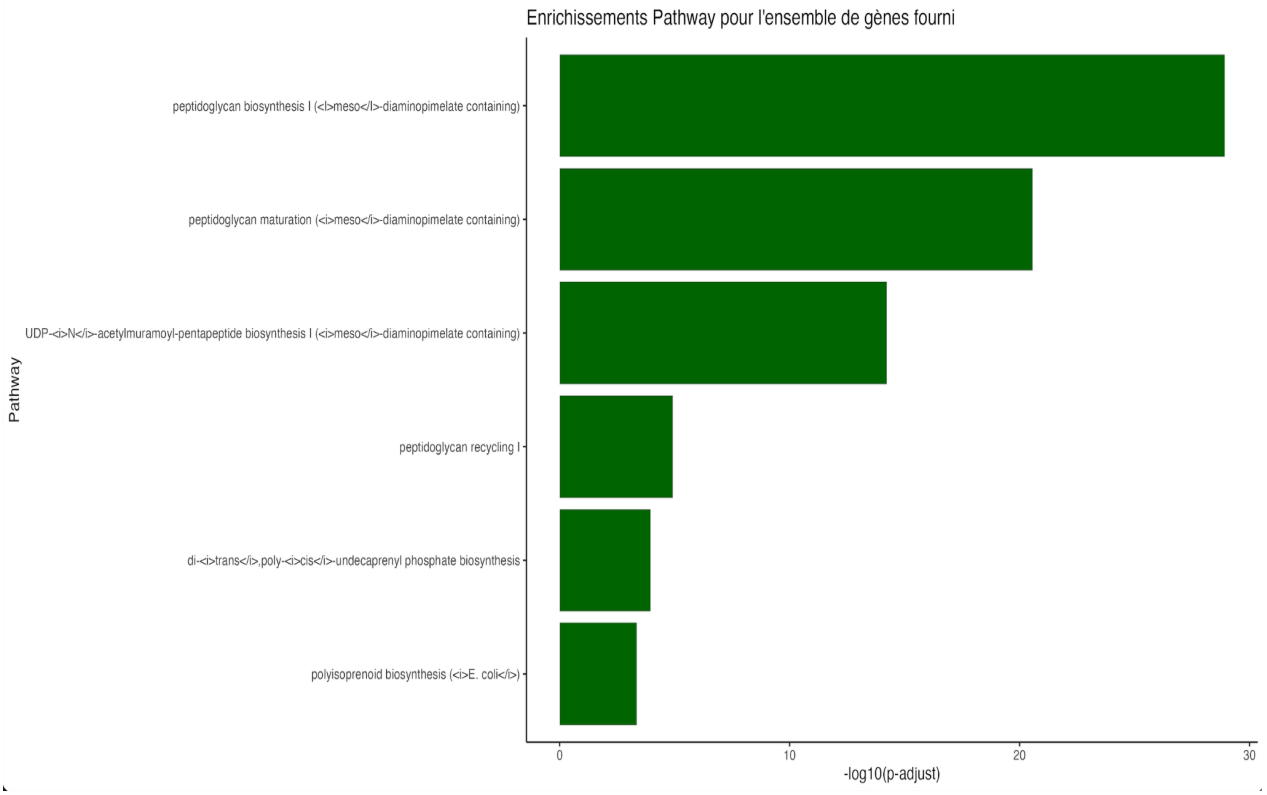

Figure 14: Enrichissements Pathway de set.02.txt

1.5 PubMed

Les publications PubMed les plus enrichies dans set.02.txt portent toutes sur la division cellulaire bactérienne, la cytokinèse et la construction du divisome. Les articles traitent notamment du rôle central de FtsZ, de la dynamique du septum, de l'ordre d'assemblage des protéines de division, et de la régulation des operons dcw et mra, impliqués dans la synthèse de la paroi et du peptidoglycane. Plusieurs publications soulignent aussi l'importance de la morphogenèse bactérienne, en cohérence avec les gènes du set.

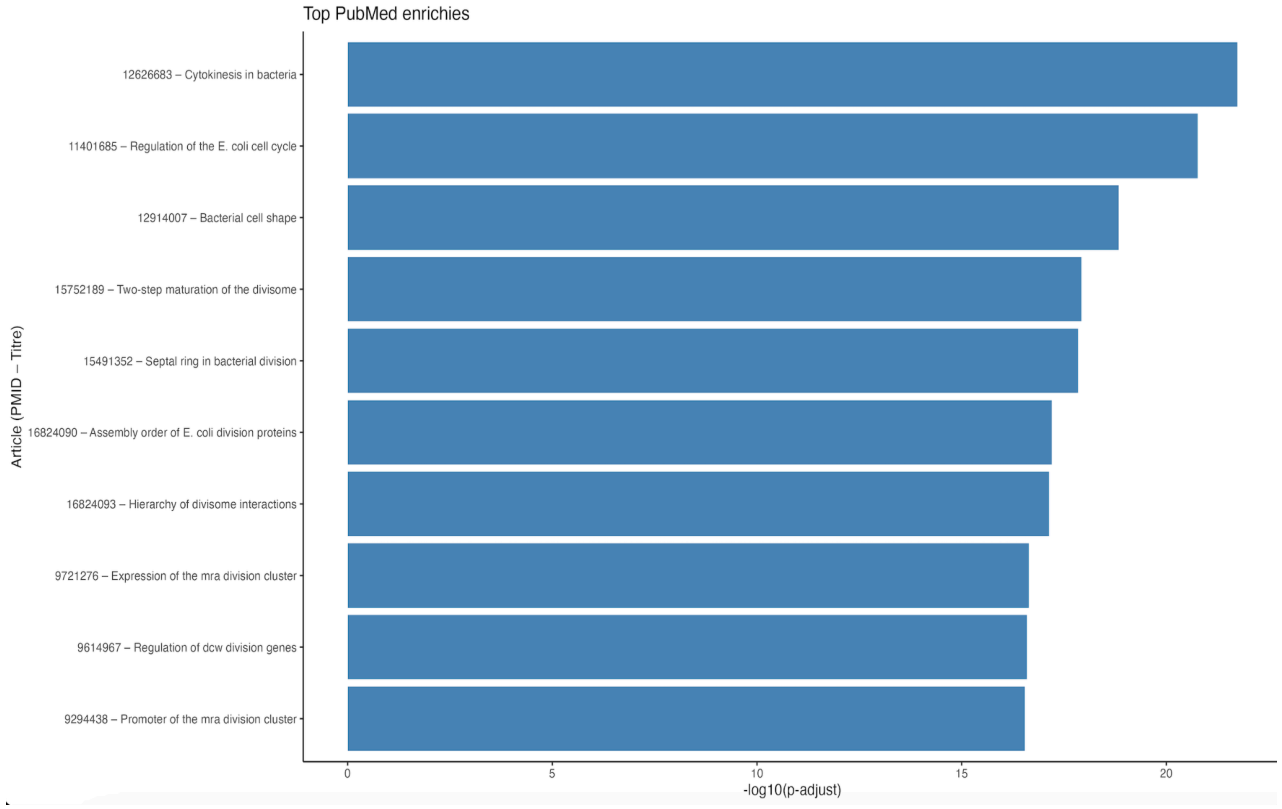

Figure 15 : Enrichissements PubMed de set.02.txt

2. Discussion et commentaires

L'ensemble set.02.txt présente une signature fonctionnelle très cohérente et centrée sur la division cellulaire bactérienne. Les différents enrichissements (GO, InterPro, Keywords, Pathways et PubMed) convergent vers un même ensemble de processus, en particulier la cytokinèse, la formation du divisome, et la biosynthèse du peptidoglycane, élément essentiel de la paroi. La présence répétée de domaines enzymatiques (transpeptidases, Mur ligases, PBPs) et de pathways associés au peptidoglycane souligne un rôle marqué dans l'assemblage et le remodelage structural de la cellule.

Les résultats montrent également une forte implication des gènes dans le maintien de la morphologie et la régulation du site de division, notamment via des mécanismes dépendants de FtsZ. Les publications enrichies confirment cette interprétation en décrivant de manière détaillée l'organisation hiérarchique du divisome et la régulation transcriptionnelle des clusters dcw et mra, tous deux essentiels à la coordination entre division cellulaire et synthèse de la paroi.

Globalement, l'ensemble set.02.txt correspond à un module génétique dédié aux étapes essentielles de la division cellulaire d'E. coli, intégrant à la fois la formation du septum, l'organisation du divisome et la synthèse du peptidoglycane.

VI. Références

Huang, D. W., Sherman, B. T., & Lempicki, R. A. (2009). *Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.*

Bilan personnel sur le projet et l'UE

Ce projet nous a permis de mieux comprendre l'intégration de données biologiques et l'analyse d'enrichissement. L'utilisation de Neo4j et la comparaison des mesures statistiques nous ont donné une vision concrète des approches fonctionnelles. Malgré quelques étapes techniques parfois longues, le projet était formateur.

Globalement, cette UE nous a beaucoup appris, autant sur le plan méthodologique que pratique, et nous a donné une meilleure compréhension du lien entre données biologiques et interprétation fonctionnelle.

Répartition du travail:

1. **Partie code:**

- **Issa:**
  - Implémentation du code R pour l'intégration des données dans la base de données Neo4j
  - Ajout de fonctionnalités au script Python de recherche d'enrichissement
  - Implémentation du script Python pour la comparaison des différentes mesures d'enrichissement
- **Hawa:**
  - Réalisation des statistiques descriptives sur les données intégrées
  - Développement du code R pour les graphes descriptifs et les figures d'enrichissement pour l'analyse visuelle permettant l'interprétation des résultats.

2. **Partie rédaction**

- **Hawa:**
  - Introduction
  - Intégration et Préparation des données
  - Analyse et interpretation des resultats d'enrichissement de set.02.txt
- **Issa:**
  - Introduction
  - Ajout de fonctionnalités
  - Comparaison des mesures intégrées
  - Analyse et interpretation des resultats d'enrichissement de set.M2.14.txt

Aide extérieure

Les supports de cours et les travaux pratiques de l'UE nous ont apporté les bases nécessaires pour comprendre les notions abordées et mettre en œuvre les différentes étapes du projet. L'article de Huang et al. (2009) a servi de référence bibliographique pour situer les méthodes d'enrichissement dans leur contexte général. Nous avons également eu recours à ChatGPT pour rectifier des erreurs d'orthographe si nécessaire.