EVOL TME 02

WANG Jinxin 3404759
Spécialité Bioinformatique et Modélisation
MASTER D'INFORMATIQUE Niveau 2
Département Master d'Informatique
Université Pierre et Marie CURIE

A : Récupérer les génomes à étudier

Une python script gb2aa.py est implémenté pour extrait CDS et traduit les gênes aux séquences protéiques.

La script prend trois paramètres obligatoires :

paramètres	*Explications*
-gb ou --geneBank-file-name	*Gene bank file name to read*
-f ou --fasta-file-name	*Fasta file name to read*
-o ou --out-fasta-file-name	*Fasta file name to write*

B : Construction des familles de gènes

Effectuer les comparaisons nécessaires pour :
Dresser la liste des Reciprocal Best Hits entre les 3 paires de génomes.

Pour réaliser le sujet, la base de donnée sont construit d'abord en utilisant makeblastdb, les données sont alignés par blastp.

Pour séléctioner les meilleurs hits, select_best_blast_hits.py as plusieurs paramètres :

paramètres	*Explications*
-f ou --hits-file-name	*input blastp hits file name*
-o ou --output-file-name	*output blastp hits file name*
-i ou --iden-percent	*% identity*
-a ou --align-len	*alignment length*
-m ou --mismat	*mismatches*
-g ou --gaps	*gap opens*
-e ou --evalue	*evalue*
-b ou --bit-score	*bit score*

J'ai choisi evalue comme la seuil à 1, par conséquence, tous les hits qui ont une evalue supérieur que 1 sont rejetés.

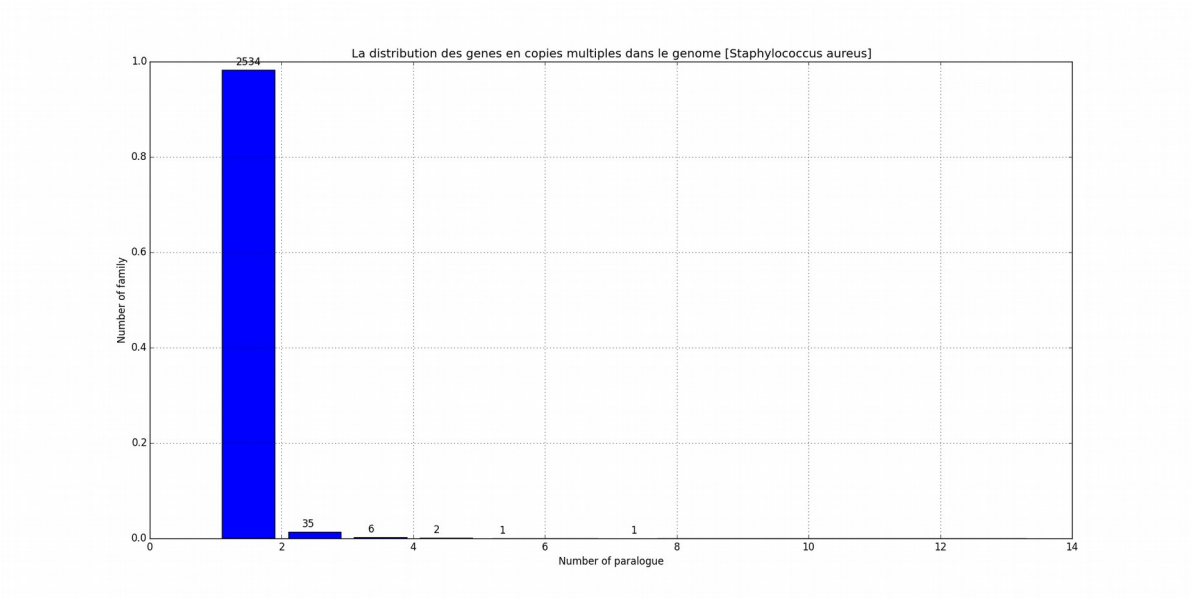
Dresser la liste des familles de gènes homologues partagés par ces 3 génomes.

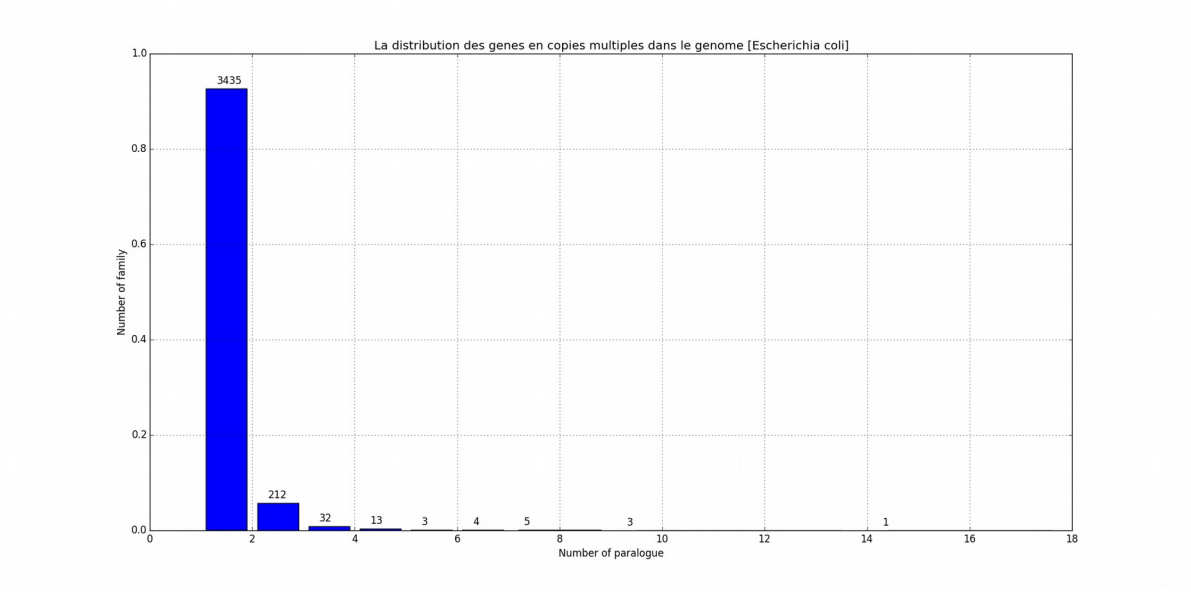
Ensuite, les paramètres par défaut de silix sont choisi.

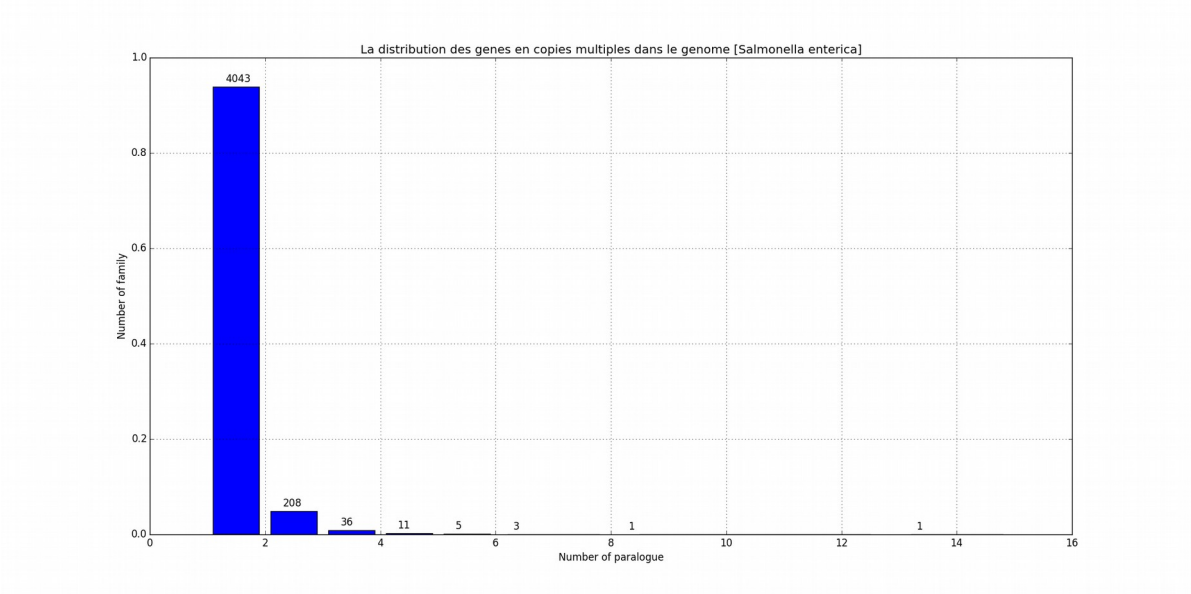
C : Analyse

- Pour chaque génome, faire un histogramme de la distribution des gènes en copies multiples dans le génome.

La script dist_hist.py est implémentée pour ploter les histogrammes.

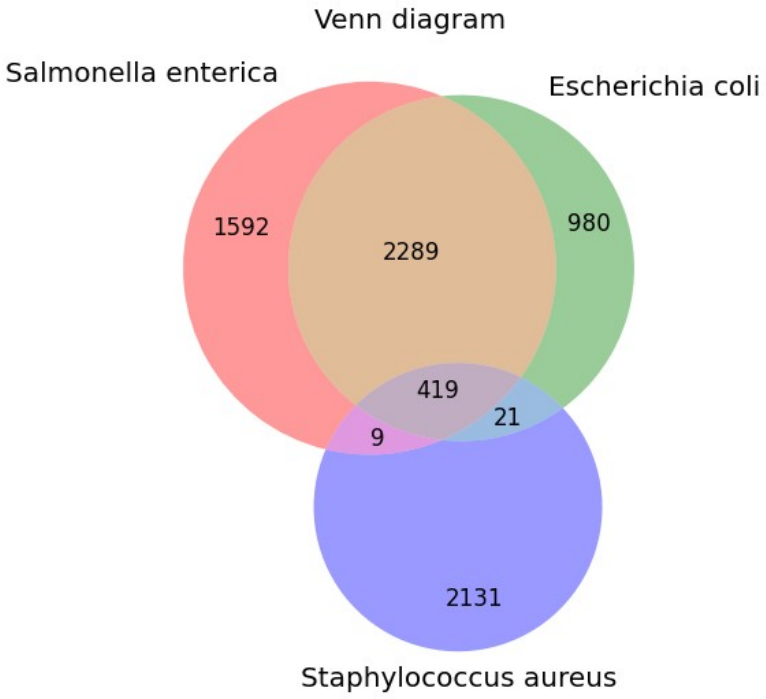






- Réaliser un diagramme de Venn afin de visualiser :

Le plot est généré par venn_list.



- Proposer une première analyse biologique de ces résultats

Dans les histogrammes de la distribution des gènes en copies multiples dans le génome, on observe que les orthologues sont très et diminuer exponentiellement. On peut aussi conclure que dans les gênes, il y a des gênes très exprimés.

Le diagramme de Venn montre que Salmonella Enterica et Escherichia Coli ont beaucoup d'orthologues en commun, mais un peu éloigné de Staphylococcus Aureus. Pour cela, nous pouvons dire que les activés des Salmonella Enterica et Escherichia Coli sont plus proches que Staphylococcus Aureus.