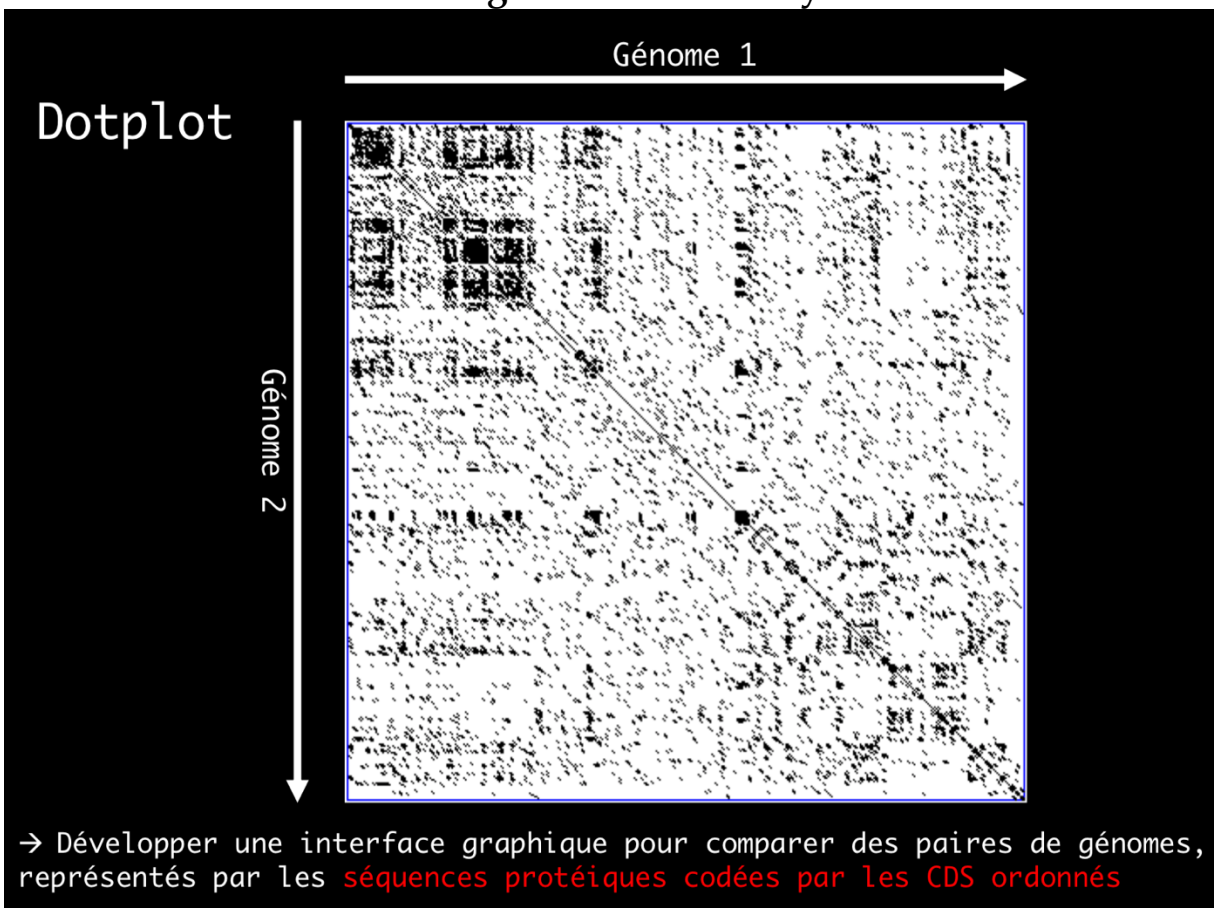
Comparaison de paires de génomes bactériens par dotplot : familles de gènes et blocs de synténie



Objectif du projet

Développer une **interface graphique** permettant, pour une **paire de génomes bactériens complets** choisie par l'utilisateur et dont les informations seront stockées dans une **base de données relationnelle**, de représenter par un dotplot les gènes homologues et leur position dans ces génomes. Cette interface devra également permettre d'identifier les **blocs de synténie**, c'est-à-dire des régions au sein desquelles l'ordre des gènes est conservé. Une discussion concernant la conservation des répertoires génomiques est attendue. Il est vivement recommandé d'illustrer cette analyse pour des paires de protéomes correspondant à des souches d'une même espèce, d'un même genre ou plus éloignées d'un point de vue taxonomique.
Vous devrez apporter un soin particulier à l'interface graphique proposée.

Des gènes pourront être considérés comme homologues selon différents critères :

- pourcentage d'identité (blast)
- E-value associée au hit (blast, COG)
- couverture du hit sur les séquences (blast)
- annotation fonctionnelle (famille ou catégorie fonctionnelle COG)

Pour les critères quantitatifs, un seuil sera fixé. Il est vivement recommandé de tester différents seuils et de comparer les résultats correspondants. Cette comparaison devra être discutée.

A consulter
https://ftp.ncbi.nlm.nih.gov/genomes/GENOME_REPORTS/
https://ftp.ncbi.nih.gov/pub/taxonomy/ (en particulier : archive *taxdump.tar.gz*)

Exemples de protéomes à comparer

- Espèce *Escherichia coli* : souche K-12 (substr. MG1655), souche IAI1
- Genre *Ramlibacter* : *Ramlibacter* sp. 5-10, *Ramlibacter tataouinensis* TTB310
- *Magnetococcus marinus* MC-1, *Magnetospirillum magneticum* AMB-1

Données fournies pour ces protéomes (Moodle)

- Sorties blastp all-against-all (paires mentionnées ci-dessus), seuil de E–value fixé à 1e-4 (1 fichier par paire de protéomes à comparer)
- Sorties CD-search (recherche dans base de données COG v3.16), seuil de E–value fixé à 1e-4 (1 fichier par protéome)
- Liste des familles COG par catégorie fonctionnelle