Mini project part 2

- **our comparative analysis approach for gene cluster profiles:**

We Propose a novel approach to compare two gene cluster profiles (we focus in case of the same cluster in 2 different datasets) and produce a score (a number between 0 and 100) to measure how distinct the clusters are. the calculation is composed from the following parameters:

1) **X (Genetic Variation Score):** This parameter evaluates the difference in abundance between various gene orders within the clusters. We utilize a formula that directly computes the absolute difference between the gene order percentages of the two profiles. For instance:

- Profile 1: [98.06%, 1.94%]
- Profile 2: [80%, 20%]

The calculation for **X** involves:

(| 98.06 - 80 | + | 1.94 - 20 |) / 2

2) **Y (Taxa Score):** This parameter assesses the dissimilarity in taxonomic parameters between the two datasets, while considering the percentages of different orders. Similar to **X**, we directly compute the absolute difference between the taxa percentages of the two profiles.

3) **Z (Environmental Score):** Here, we quantify the discrepancy in environmental parameters between the two datasets, while also factoring in the percentages of different orders. Like **X** and **Y**, the calculation involves directly computing the absolute difference between the environmental percentages of the two profiles.

The **final_score** is computed as the average of the three individual scores(X,Y,Z)

This approach provides a straightforward assessment of the dissimilarity between gene cluster profiles, focusing solely on the absolute difference. The final score facilitates comparative analysis and aids in identifying clusters with notable distinctions across datasets.

The code initializes three dictionaries: **various_gene_orders**, **taxa_info**, and **env_info**, which are used to store information collected from part 1 code, in easy way to calculate the final score.
create_cluster_profile, **get_gene_orders_dict**, **get_taxa_info**, and **get_env_info** functions are responsible for populating the global dictionaries (**various_gene_orders**, **taxa_info**, **env_info**) with data extracted from input files (**cog_words_plasmid.txt**, **clean_file.txt**).

⌘ Comparative Gene Cluster Analysis Pipeline:

The function **find_GC(s1,s2,d,q1,q2)**- the pipeline is designed to conduct comparative analysis of gene clusters across chromosome and plasmid datasets, with the goal of identifying clusters that “differ significantly” in their profile.

1)Data Preprocessing:

Initialize different dictionaries for the data about S1 and S2.

Parsing data from separate files representing chromosome and plasmid datasets ("clean_file.txt", cog_words_plasmid.txt").

Key genomic information such as gene identifiers, COG assignments, taxa, and habitat preferences are extracted for each genome in the datasets.

2)Cluster Identification:

Use the clustering algorithm **find_clusters** based on input-specified parameters (s, d, q) from part 1. First we conducted this process for the chromosome while using S1, q1, and then we did the same for the plasmid while using S2, q2.

3)Comparative Analysis:

For each cluster that is found in both S1 and S2 it calculate the score.

If the score is above the threshold 60, then the clusters are identified as exhibiting significant profile differences.

The pipeline returns the found cluster with the given score.

⌘ **Process the data about the found cluster**

The cluster is: {'1898', '1088', '1209', '1091'}

(d: 4, q1: 1, q2: 10)

cog information:

COG	Functional Category	Role	Enzyme
COG1898	Cellular Processes and Signaling	Cell wall/membrane/envelope biogenesis	dTDP-4-dehydrorhamnose 3,5-epimerase or related enzyme
COG1088	Cellular Processes and Signaling	Cell wall/membrane/envelope biogenesis	dTDP-D-glucose 4,6-dehydratase
COG1209	Cellular Processes and Signaling	Cell wall/membrane/envelope biogenesis	dTDP-glucose pyrophosphorylase
COG1091	Cellular Processes and Signaling	Cell wall/membrane/envelope biogenesis	dTDP-4-dehydrorhamnose reductase

These enzymes are interconnected through their involvement in common metabolic pathways and the shared substrates and products in their respective reactions. Let's break down the connections:

1**. COG1898: dTDP-4-dehydrorhamnose 3,5-epimerase**

- Reaction: dTDP-4-dehydro-6-deoxy-D-glucose ⇌ dTDP-4-dehydro-6-deoxy-L-mannose

2. **COG1088: dTDP-glucose 4,6-dehydratase**

- Reaction: dTDP-glucose ⇌ dTDP-4-dehydro-6-deoxy-D-glucose + H2O

- Connection to COG1898: COG1088 produces a substrate (dTDP-4-dehydro-6-deoxy-D-glucose) for COG1898.

3**. COG1209: Glucose-1-phosphate thymidylyltransferase**

- Reaction: dTTP + alpha-D-glucose 1-phosphate ⇌ diphosphate + dTDP-glucose

- Family: Transferases (nucleotidyltransferases)

- Pathways: Nucleotide sugars metabolism, streptomycin biosynthesis, polyketide sugar unit biosynthesis.

- Connection to COG1088: COG1209 produces dTDP-glucose, a substrate for COG1088.

4. **COG1091: dTDP-4-dehydrorhamnose reductase**

- Reaction: dTDP-6-deoxy-L-mannose + NADP+ ⇌ dTDP-4-dehydro-6-deoxy-L-mannose + NADPH + H+

- Connection to COG1898: COG1091 shares a product (dTDP-4-dehydro-6-deoxy-L-mannose) with COG1898.

Category	Plasmid (%)	Chromosome (%)
1. order: 1088120918981091	8%, 1	23%, 4
Proteobacteria	8%, 1	21%, 4
Not Annotated	8%, 1	3%, 1
Marine Environment	0%, 0	3%, 1
Soil and Sediment	0%, 0	6%, 1
Plant	0%, 0	7%, 1
2. order: 1898108812091091	8%, 1	1%, 0
Acidobacteria	8%, 1	0%, 0
Not Annotated	8%, 1	0%, 0
3. order: 1898108810911209	46%, 6	14%, 2
Proteobacteria	46%, 6	14%, 2
Plant	8%, 1	5%, 1
Not Annotated	15%, 2	1%, 0
Soil and Sediment	8%, 1	4%, 1
Animal	8%, 1	0%, 0
Marine Environment	8%, 1	1%, 0
4. order: 1898109112091088	8%, 1	2%, 0
Spirochaetes	8%, 1	0%, 0
Animal	8%, 1	0%, 0
5. order: 1209189810911088	8%, 1	6%, 1
Fusobacteria	8%, 1	0%, 0
Proteobacteria	0%, 0	4%, 1
Soil and Sediment	8%, 1	3%, 1
6. order: 1088109118981209	8%, 1	0%, 0
Firmicutes	8%, 1	0%, 0
Host	8%, 1	0%, 0
7. order: 1209189810881091	15%, 2	22%, 4
Proteobacteria	15%, 2	3%, 1
Firmicutes	0%, 0	19%, 3
Soil and Sediment	8%, 1	3%, 1
Host	8%, 1	7%, 1
Human	0%, 0	3%, 1
Animal	0%, 0	4%, 1
8. order: 1088109112091898	0%, 0	23%, 4
Proteobacteria	0%, 0	23%, 4
Soil and Sediment	0%, 0	5%, 1
Not Annotated	0%, 0	3%, 1
Host	0%, 0	6%, 1
Fresh water	0%, 0	3%, 1
9. order: 1088120910911898	0%, 0	6%, 1
Proteobacteria	0%, 0	6%, 1

The genes RfbD, RfbB, RfbC, and RmlA1, identified in the study : ***O-Antigen-Dependent Colicin Insensitivity of Uropathogenic Escherichia coli***.

These play roles in the biosynthesis of lipopolysaccharide (LPS), a crucial component of the outer membrane of Gram-negative bacteria. These genes are involved in the O-antigen synthesis pathway.

The O-antigen is the outermost component of the LPS molecule and contributes to the structural and functional properties of the bacterial outer membrane. The study suggests that the density of O-antigen is a critical factor in determining the sensitivity of uropathogenic Escherichia coli (UPEC) to colicins.

The density of the O-antigen, influenced by the activities of these genes, affects the ability of colicins to bind to their receptors and, consequently, the sensitivity of the bacteria to colicin-mediated killing.

After observation of the cluster profile we found, considering the different orders, habitats, and taxa information, there are some findings.

There are some data we found only in the plasmid:

- (order number 2 ,1, Acidobacteria, Not Annotated)
- (order number 3 ,1, Proteobacteria, Animal)
- (order number 4, 1, Spirochaetes, Animal)
- (order number 6,1, Firmicutes, Host)

We noticed different distributions of the factors we tested, and it also appears that the gene cluster is more abundant in plasmid relative to chromosomes. Therefore, we suggest several hypotheses of advantages that clusters give in relation to the article and the role of the genes in the biosynthesis of lipopolysaccharides (LPS):

1. **Functionality in Variable Environments:** The prevalence of LPS biosynthesis genes on plasmids, observed across various habitats (e.g. Acidobacteria, Proteobacteria, Spirochaetes, and Firmicutes), suggests an adaptive advantage for bacteria in fluctuating environments. Plasmids enable rapid modulation of cell surface structures, crucial for survival in diverse conditions.
2. **Horizontal Gene Transfer (HGT):** The consistent presence of these genes on plasmids indicates their potential for horizontal gene transfer (HGT) among bacterial populations. This genetic exchange allows for the acquisition of traits related to outer membrane composition and colicin insensitivity across different taxa.
3. **Ecological Niche Adaptation:** The data showing plasmid localization of LPS biosynthesis genes in taxa associated with specific ecological niches (e.g. Animal) suggests their role in niche adaptation. This localization likely enhances bacterial fitness within these niches, possibly through evasion of host immune responses.
4. **Selective Pressure and Fitness Cost:** Despite the fitness cost associated with plasmid maintenance, the widespread distribution of

these genes on plasmids indicates that the benefits outweigh the costs. The selective pressures imposed by diverse environments likely drive the retention of these genes on mobile genetic elements.

In summary, the abundance of the gene cluster {1898, 1088, 1209, 1091} in plasmids compared to chromosomes suggests that these genes provide selective advantages to bacteria when located on plasmids.

---

related article**: *O-Antigen-Dependent Colicin Insensitivity of Uropathogenic Escherichia coli***

Citation: *Sharp C, Boinett C, Cain AHousden NG, Kumar S, Turner K, Parkhill J, Kleanthous C.2019.O-Antigen-Dependent Colicin Insensitivity of Uropathogenic Escherichia coli. J Bacteriol201:10.1128/jb.00545-18.https://doi.org/10.1128/jb.00545-18*

Summary:

The article discusses the phenomenon of colicin insensitivity in Escherichia coli (E. coli) strains, particularly focusing on uropathogenic E. coli (UPEC) sequence type 131 (ST131). Colicins are antimicrobial proteins secreted by E. coli, and they have the potential to be used as narrow-spectrum antimicrobials.

The researchers demonstrate that this insensitivity can be overcome by making minor changes to the environmental conditions.

They found that the production of O-antigen is a key factor in colicin insensitivity but that slight modifications to growth conditions can render the organism sensitive to colicins. The reintroduction of O-antigen into E. coli K-12 demonstrated that the density of O-antigen is a dominant factor governing colicin insensitivity.

One of the measures that was used in the research: A transposon-directed insertion sequencing (TraDIS) approach was employed to identify genes associated with colicin insensitivity in UPEC ST131. Thirty-three genes (four of them are the genes in the cluster we found) were mapped to operons responsible for lipopolysaccharide (LPS) biosynthesis, O-antigen biosynthesis, and other outer membrane-related processes.

Fluorescence microscopy revealed that the presence of O-antigen diminished colicin binding to the outer membrane receptor (BtuB). Changes in O-antigen density influenced colicin binding and sensitivity.

In summary, the study provides insights into the mechanisms and environmental factors influencing colicin insensitivity in UPEC ST131, shedding light on the complex interplay between bacterial genetics, outer membrane composition, and environmental conditions.