CHALLENGE DATA 2023-2024

23 JANVIER 2024

3A-IT _ 41004795
GAIDO—AMOROS Alexia

ESIROI
ÉCOLE D'INGÉNIEURS
UNIVERSITÉ DE LA RÉUNION

Présentation de la base de données

La base de données utilisée dans le cadre de cette analyse constitue le fondement de notre système d'analyse des distributions d'acides aminés.

Structure de la Base de Données

La base de données, nommée "database", abrite une collection principale appelée "freq_taxon". Chaque document au sein de cette collection contient des informations sur la fréquence des acides aminés pour divers taxons.

La structure de chaque document dans la collection "freq_taxon" est définie comme suit :

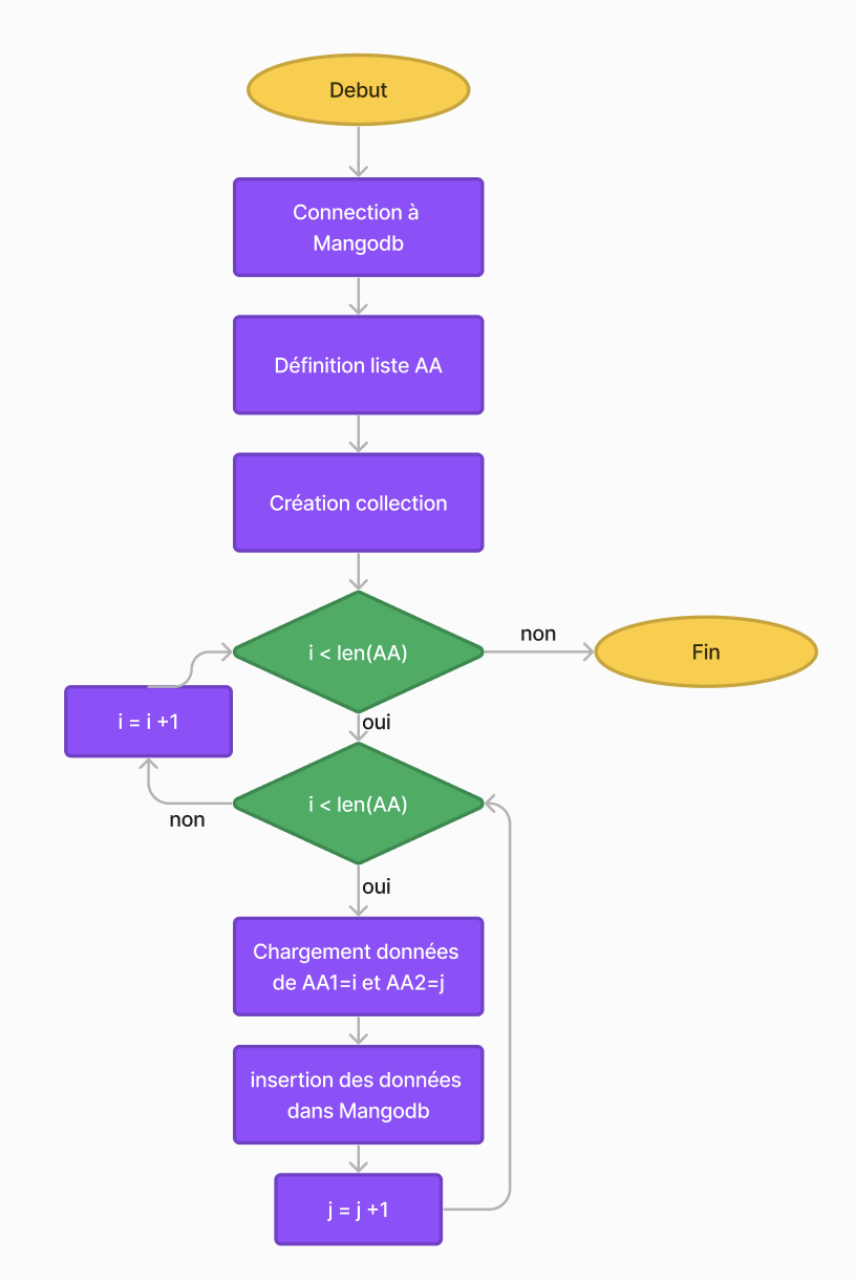
_id : Identifiant unique du document, généré automatiquement par MongoDB.
AA1 et AA2 : Acides aminés analysés dans la distribution.
taxon_id : Identifiant du taxon associé à la fréquence.
Count : Nombre total d'occurrences du motif d'acides aminés pour le taxon donné.
Colonnes de 1 à 100 : Représentent la position dans la séquence des acides aminés, avec les valeurs de fréquence correspondantes.

Les acides aminés (AA) sont représentés par des codes à une lettre, notamment : A (Alanine), C (Cystéine), D (Aspartate), E (Glutamate), F (Phénylalanine), G (Glycine), H (Histidine), I (Isoleucine), K (Lysine), L (Leucine), M (Méthionine), N (Asparagine), P (Proline), Q (Glutamine), R (Arginine), S (Sérine), T (Thréonine), V (Valine), W (Tryptophane), et Y (Tyrosine).

Infrastructure et Hébergement

L'infrastructure de la base de données repose sur une instance MongoDB déployée sur le serveur vps1.amoros.io. Cette décision a été prise en raison de problèmes techniques rencontrés sur ma machine locale, nécessitant un environnement plus robuste.

Logigramme pour l’insertion des données en base



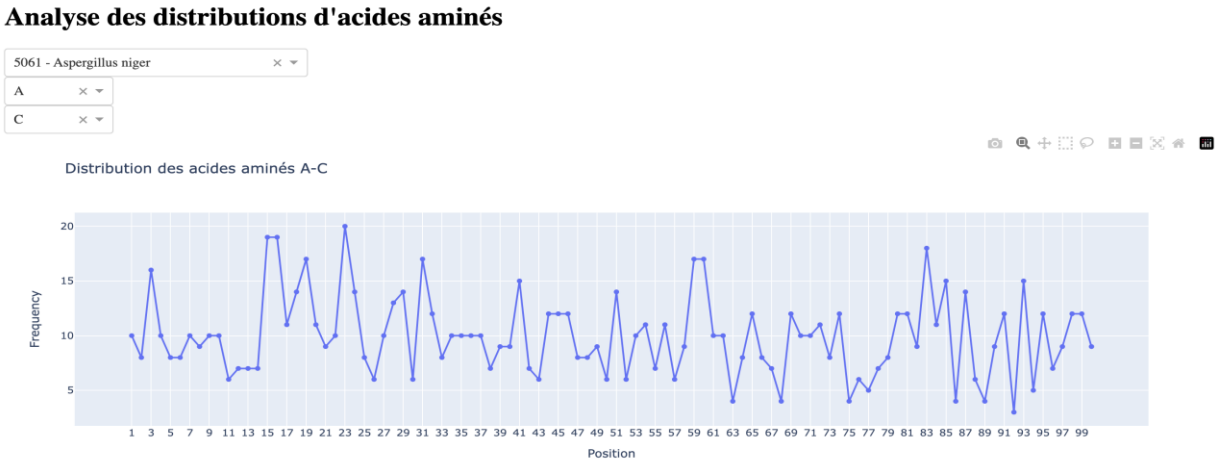
Visualisations réalisées

Graphique de Distribution des Acides Aminés :

Ce graphique de ligne montre comment la distribution des acides aminés varie le long de la séquence d'acides aminés (positions 1 à 100). La ligne représente la

distribution d'une paire spécifique d'acides aminés (AA1 et AA2) pour une espèce sélectionnée.

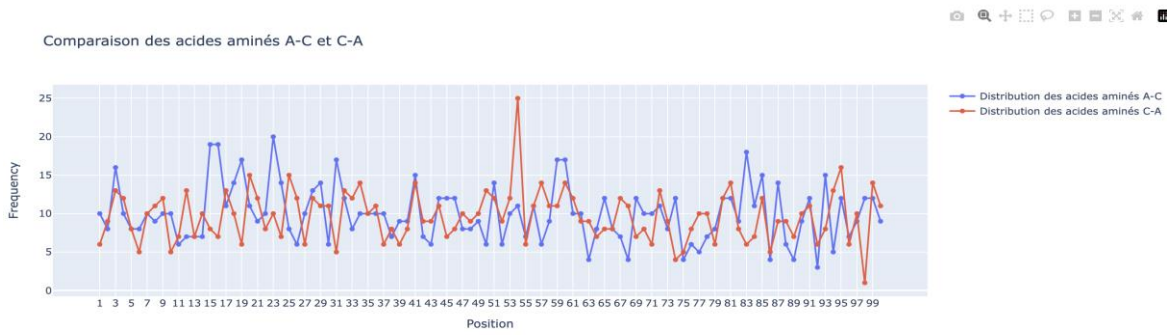
Ce graph permet d'identifier les positions spécifiques où certains acides aminés apparaissent fréquemment ou rarement. Cela peut être utile pour détecter des motifs ou des variations significatives dans la composition des acides aminés pour une espèce donnée.



Graphique de Comparaison des Acides Aminés :

Ce graphique de ligne compare la distribution des acides aminés pour deux paires différentes (AA1-AA2 et AA2-AA1) sur la même séquence d'acides aminés.
Chaque ligne représente la distribution d'une paire spécifique d'acides aminés pour une espèce sélectionnée.

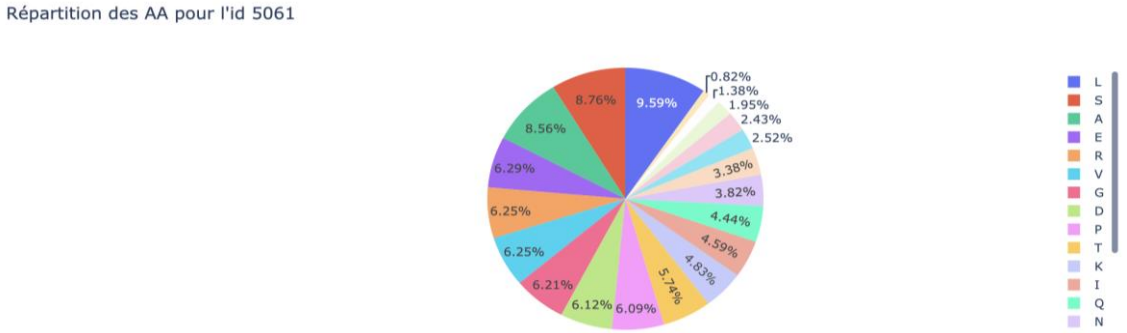
Cela peut faciliter la comparaison visuelle entre deux paires d'acides aminés. Cela peut aider à identifier des différences significatives ou des similitudes dans la distribution des acides aminés entre les deux paires, contribuant ainsi à l'analyse comparative.



Graphique de Répartition des Acides Aminés :

Ce diagramme circulaire (pie chart) représente la répartition des occurrences des acides aminés (A, C, D, ..., Y) pour une espèce sélectionnée. Chaque segment du cercle correspond à un acide aminé spécifique.

Cela fournit une vue d'ensemble visuelle de la fréquence relative de chaque acide aminé dans la séquence pour une espèce donnée. Cela peut être utile pour identifier les acides aminés dominants ou inhabituels dans la composition protéique de l'espèce.



Perspectives d’extension du travail

L'introduction d'une analyse comparative entre espèces et la création de graphiques de comparaison entre parents et enfants amélioreraient l'étude sur la distribution d'acides aminés.

L'Analyse Comparative entre Espèces permettrait de contextualiser les distributions d'acides aminés en les replaçant dans un cadre évolutif. En comparant les profils entre différentes espèces, cette approche offrirait des informations sur l'évolution des caractéristiques moléculaires, élargissant la portée de l'analyse.

L'Ajout de Graphiques de Comparaison Parents-Enfants permettait quant à elle d'explorer l'hérédité des distributions d'acides aminés. Les graphiques pourraient révéler des modèles de transmission génétique ou des variations notables entre générations, offrant un aperçu des mécanismes de transmission des caractéristiques moléculaires.

Ces améliorations répondent à des questions plus complexes sur l'évolution moléculaire et l'hérédité, ajoutant ainsi de la profondeur à l'analyse.

Lien vers le dépôt et ressources utilisé

Ressources utilisées :
https://pymongo.readthedocs.io
https://www.youtube.com/@CharmingData
https://plotly.com/python/time-series/
(Ainsi que d'autre que je oublié d'enregistrer)

Lien vers le git :
https://github.com/Aralta/ChallengeData