TELECOM nancy
Ingénieurs du numérique • Inspiring your digital future

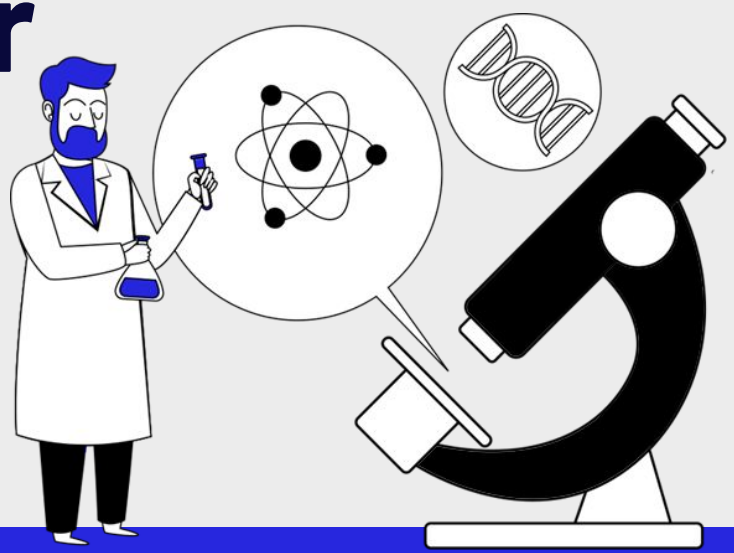
UNIVERSITÉ DE LORRAINE

Protein Analyzer

Projet NoSQL - IAMD

Maxence Agra - Lucine Giraud - Lina Lekbouri

17 décembre 2025



Datasets utilisés

→ Données UniProt sur les souris et les humains

`uniprot-compressed_true_download_true_fields_accession_2Cid_2Cprotei-2022.11.14-07.52.02.48.tsv`
`uniprotkb_AND_model_organism_10090_2025_11_14.tsv`

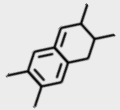
Docker

→ service **mongo**
→ service **neo4j**
→ service **app** : crée à part d'un Dockerfile personnalisé → lance les scripts de création des bases de données mongo et neo4j

Tâche 1 : Stockage documentaire avec MongoDB

Stockage sous forme de **document JSON**

⇒ Permet une recherche textuelle et un filtrage via des requêtes uniques

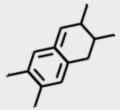


Tâche 1 : Stockage documentaire avec MongoDB

Stockage sous forme de **document JSON**

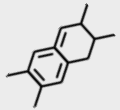
⇒ Permet une recherche textuelle et un filtrage via des requêtes uniques

```
{
  "_id": "A0A024QYR9",                    // = Entry (UniProt ID)
  "uniprot_id": "A0A024QYR9",
  "entry_name": "A0A024QYR9_MOUSE",       // Entry Name
  "organism":  …,                           // if present
  "protein_names": [
    "Phosphatidylinositol",
    "3,4,5-trisphosphate", …
  ],
  "sequence": {
    "length": 572,   //Calcul à faire
    "aa": "MERGGEAAAAAAAPGRGSESPVTI..."     // Sequence brute
  },
  "interpro_ids": [
    "IPR035892",
    "IPR051281", …
  ],
  "ec_numbers": [
    "3.1.3.16",
    "3.1.3.48"...
  ],
  "is_labelled": true           // true si au moins un EC present
}
```



✅ **Tâche 1 :** Stockage documentaire avec MongoDB

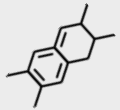
Tâche 2 : Construction du graphe avec neo4j



✅ Tâche 1 : Stockage documentaire avec MongoDB

Tâche 2 : Construction du graphe avec neo4j

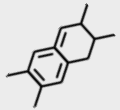
- Méthode : Nœuds : Protéines reliées par leurs Domaines communs.
- Algorithme : Indice de Jaccard via Graph Data Science (GDS).
- Poids : Force du lien basée sur le % de similarité (Intersection / Union).



✅ **Tâche 1 :** Stockage documentaire avec MongoDB

✅ **Tâche 2 :** Construction du graphe avec neo4j

Tâche 3 : Interroger les bases de données



✅ Tâche 1 : Stockage documentaire avec MongoDB

✅ Tâche 2 : Construction du graphe avec neo4j

Tâche 3 : Interroger les bases de données

Flask
web development, one drop at a time

Classes manager pour chaque base de données :
`MongoProteinQueryManager` et **`Neo4jProteinQueryManager`**

Utilisée par notre API Flask

→ Recherche textuelle, recherche des voisins dans le graphe, statistiques…
→ Visualisation de graphe avec Cytoscape

Cytoscape

✅ **Tâche 1 :** Stockage documentaire avec MongoDB

✅ **Tâche 2 :** Construction du graphe avec neo4j

✅ **Tâche 3 :** Interroger les bases de données

Tâche 4 : Annotation fonctionnelle

✅ **Tâche 1 :** Stockage documentaire avec MongoDB

✅ **Tâche 2 :** Construction du graphe avec neo4j

✅ **Tâche 3 :** Interroger les bases de données Flask (web development, one drop at a time)

Tâche 4 : Annotation fonctionnelle

3 étapes :

- Calcul des communautés de protéines
- Choix d'un algorithme d'écriture des ECs
  - Vote majoritaire
  - Union
- Application de l'algorithme

✅ **Tâche 1 :** Stockage documentaire avec MongoDB

✅ **Tâche 2 :** Construction du graphe avec neo4j

✅ **Tâche 3 :** Interroger les bases de données

✅ **Tâche 4 :** Annotation fonctionnelle

Conclusion - Nombre de protéines annotées encore faible.
Améliorations :
– ajout d'autres données liées aux fonctions des protéines,
– couplage à d'autres méthodes de classification, de prédiction, de calcul de communautés.