LIP6

PLDAC M1 2025

Méthodes big data pour l'analyse de très grands graphes de protéines

Nombre d'étudiants : un binôme

Encadrant :

BD LIP6 : Hubert Naacke : Hubert.Naacke@lip6.fr ,

Partenaires du MNHN :

Mathilde Carpentier Mathilde.Carpentier@mnhn.fr et Lucie Bittner Lucie.Bittner@sorbonne-universite.fr

Contexte

Ce projet s'intéresse à des grandes bases de données issues du séquençage de protéines. Ces données sont produites et analysées par l'atelier de Bio-Informatique (Muséum National d'Histoire Naturelle). L'objectif général est de développer des méthodes pour exploiter des ensembles séquences protéiques. Dans ce contexte, un **graphe de similarité** entre des séquences protéiques a été produit. Un nœud de ce graphe est un identifiant de protéine et un arc relie deux protéines en précisant le score de similarité (allant de 80% à 100% d'identité) et la longueur de la sous-séquence commune. On connaît aussi, pour certaines protéines, leurs propriétés fonctionnelles décrites par une liste de labels. Ces annotations fonctionnelles sont incomplètes et la plupart des protéines ne sont pas annotées.

Ce projet aborde essentiellement les défis liés à la très grande volumétrie des données : les graphes de similarité à analyser contiennent plus de 20 milliards d'arcs et limitent l'utilisation des outils d'analyse existants qui nécessitent de charger les données en mémoire. Par ailleurs, les plateformes distribuées sur plusieurs machines dédiées à l'analyse de données à grande échelle ne sont pas efficaces pour exécuter les analyses envisagées qui impliquent des échanges massifs de données entre les nœuds de calcul (scalabilité limitée).

Objectif

L'objectif du projet est donc de développer ou d'adapter des méthodes d'analyse capables de traiter des graphes de protéines massifs pour la détection de sous graphes denses appelés communautés

Travail à effectuer

Etape 1 : Détection de communautés à large échelle

- Etudier l'état de l'art sur la détection de communautés et considérant deux familles d'approches : d'une part le clustering de graphe (par exemple l'algorithme de Leiden [1]) et d'autre part l'approche consistant à représenter les noeuds du graphe par des embeddings (par exemple Node2Vec [2] ou GraphSage [4]) afin d'appliquer du clustering basé sur la similarité cosinus entre les embeddings (HDBSCAN). Expliquer quelles sont leurs opportunités et limites pour le cas d'usage considéré.
- Expliciter les propriétés qu'une communauté doit satisfaire et les justifier vis-à-vis des besoins applicatifs.
- Proposer une méthode de calcul parallèle pour détecter les communautés qui satisfont les propriétés définies ci-dessus. En particulier, pour faire face à la très grande taille des données, considérer une approche de type *divide and conquer* : commencer par calculer les communautés dans une petite partie du graphe, puis compléter le calcul sur une partie de plus en plus grande du graphe. Les différentes parties du graphe pourront être définies en fonction du poids des arcs.

Etape 2 : Validation expérimentale

- Comparaison avec les solutions de l'état de l'art, en particulier avec celle implantée dans GraphX [3]. Mesurer le gain en performance de la solution proposée pour le calcul des communautés

Références

[1] The Leiden algorithm https://en.wikipedia.org/wiki/Leiden_algorithm

[2] Apac Aditya Grover, Jure Leskovec and Vid Kocijan. Node2vec: Scalable Feature Learning for Networks. ACM International Conference on Knowledge Discovery and Data Mining (KDD), 2016.

[3] Spark GraphX https://spark.apache.org/graphx/

[4] GraphSAGE: Inductive Representation Learning on Large Graphs. W.L. Hamilton, R. Ying, and J. Leskovec arXiv:1706.02216 [cs.SI], 2017. https://snap.stanford.edu/graphsage/