Stage M2 2025

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

Lieu : LIP6, 4 place Jussieu, 75005 Paris

Date : début du stage en février ou mars 2025

Encadrants :

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

ABI, MNHN : Mathilde Carpentier mathilde.carpentier@mnhn.fr, Lucie Bittner lucie.bittner@upmc.fr

Contexte

Ce stage s'inscrit dans une collaboration entre l'Atelier de Bio-Informatique du MNHN et l'équipe Bases de Données du LIP6 (UFR 919). Les progrès en séquençage de ces dernières années ont permis le séquençage de très nombreux organismes ou échantillons provenant de l'environnement, mais l'analyse de ces masses de données est difficile à cause de la quantité de données et de la difficulté d'avoir des données expérimentales. L'objectif général est ici de développer des méthodes pour permettre d'exploiter les séquences protéiques.

Dans ce contexte, et dans le but d'annoter des séquences protéiques, des algorithmes de propagation de labels ont été proposés pour estimer les labels manquants en s'appuyant sur des représentations sous forme de **graphes de similarité** entre des séquences protéiques. Dans ce graphe, un nœud 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 stage 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 du stage

L'objectif du stage est donc de développer ou d'adapter des méthodes d'analyse capables de traiter des graphes de protéines massifs pour (1) la détection de sous graphes denses appelés communautés et (2) la prédiction et la fusion de label pour les protéines sans label.

Travail à effectuer

Le stage se déroulera en plusieurs étapes afin de mieux séparer les défis à aborder concernant la prise en compte de la très grande taille d'un graphe (étape 1), l'analyse prédictive dans un graphe (étape 2).

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 : Prédiction et fusion de labels

- Prédiction de labels : On suppose connu le label (*i.e.* l'annotation fonctionnelle) de certains nœuds. A partir des labels connus dans chaque composante, proposer une méthode permettant d'attribuer un label aux nœuds sans label. Définir le taux de confiance pouvant être accordé à l'attribution d'un label.
- Fusion de labels : Étudier la cooccurrence entre les labels. La cooccurrence est le nombre de composantes contenant les deux labels. Proposer une méthode basée sur cette cooccurrence pour repérer des petits ensembles de labels qui co-occurrent souvent entre eux mais rarement avec d'autres labels. Puis déterminer quels labels pourraient être fusionnés et avec quelle certitude.

Etape 3 : 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 à la fois pour le calcul des communautés et pour la prédiction de labels.

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/