Clustering de protéines et inférence de familles fonctionnelles

PMIND 2026

Encadrants

Hubert Naacke (hubert.naacke@lip6.fr) et Sara Jarrad (sara.jarrad@lip6.fr)

Mots clés

Gestion de grandes bases de données de protéines, clustering de graphes, représentation vectorielle (embedding) de protéines, prediction d'annotation.

Contexte

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 d'exploiter les séquences protéiques afin de déterminer (inférer) des groupes fonctionnels de protéines. La tâche visée consiste à prédire les fonctions d'une protéine. Une difficulté est que l'information dont on dispose est très partielle : les fonctions sont connues seulement pour moins de 25% des protéines. Les données étudiées sont :

- Réseau de similarité *Blast* : données d'alignement entre les paires de protéines, calculé par la méthode Blast.
- Embedding d'une protéine. C'est un vecteur dense qui capture des relations complexes provenant de la structure interne d'une protéine. Ces embeddings ont été obtenus à partir de modèles pré-entrainés [1].
- Annotations fonctionnelles : certaines protéines ont une ou plusieurs annotations fonctionnelles déterminée. On dispose aussi d'information sur l'incompatibilité entre annotations fonctionnelles. Deux annotations incompatibles ne doivent pas coexister dans un même cluster de protéines.

L'objectif du projet est d'améliorer le processus de clustering en exploitant, en complément au réseau de similarité Blast, les embeddings des protéines.

Défis : les méthodes de clustering de graphe existantes [2, 3] ne donnent pas des bons résultats lorsque qu'elles sont appliquées sur un réseau de similarité de protéines. Les clusters obtenus ne sont pas assez *homogènes* : certains clusters contiennent des protéines ayant des annotations fonctionnelles incompatibles. Le **résultat attendu** est une méthode permettant de détecter des clusters représentant des groupes plus homogènes.

Méthodes

Après avoir défini formellement le score d'homogénéité d'un cluster, Plus précisément, il s'agit de proposer une **méthode de clustering** basée sur la similarité Blast **et** sur les embeddings des protéines qui permette d'obtenir des clusters plus homogènes.

Plusieurs approches seront envisagées telles que : (1) *Aligner* les clusters obtenus à partir du réseau de similarité et ceux obtenus à partir des embeddings. (2) modifier les *poids* dans le graphe de similarité en combinant la similarité entre séquences avec la similarité cosinus entre leurs embeddings.

Les approches seront validées expérimentalement pour comparer leur efficacité (score d'homogénéité) et leurs performances (temps de calcul en fonction de la taille du graphe).

Références

[1] Nadav Brandes, Dan Ofer, Yam Peleg, Nadav Rappoport, and Michal Linial. ProteinBERT : a universal deep-learning model of protein sequence and function. *Bioinformatics*, 38(8) :2102–2110, February 2022. _eprint : https ://academic.oup.com/bioinformatics/article-pdf/38/8/2102/49009610/btac020.pdf.

[2] Vincent A. Traag, Ludo Waltman, and Nees Jan van Eck. From louvain to leiden : guaranteeing well-connected communities. *CoRR*, abs/1810.08473, 2018.

[3] Stijn Van Dongen and Cei Abreu-Goodger. Using mcl to extract clusters from networks. In *Bacterial molecular networks : Methods and protocols*, pages 281–295. Springer, 2011.