

Annotation fonctionnelle de protéines par intégration de graphes hétérogènes massifs et apprentissage multimodal à large échelle

Hubert Naacke, équipe BD, LIP6

en co-encadrement avec Mathilde Carpentier et Lucie Bittner, équipe ABI, MNHN

mars 2026

Mots-clés : Graphes hétérogènes multimodaux, Prédiction de fonctions protéiques, Apprentissage de représentations, Passage à l'échelle.

Contexte scientifique et problématique

L'explosion du séquençage produit un volume considérable de séquences protéiques dont le nombre atteint maintenant plusieurs milliards¹. Ces séquences sont de plus issues d'organismes non modèles, révélant ainsi une diversité moléculaire encore largement inexplorée. Un enjeu majeur pour la biologie évolutive est d'analyser ces données afin de mieux comprendre la fonction et l'évolution des protéines. Nous nous intéressons en particulier à la prédiction de la fonction d'une protéine, sachant que la fonction de certaines protéines de référence est connue. Une méthode appelée *transfert d'annotation* est actuellement très utilisée. Elle s'appuie principalement sur la notion d'alignement, qui permet de quantifier la similarité entre deux séquences protéiques. En partant du principe que deux protéines similaires partagent très probablement une origine évolutive commune et la même fonction, on peut attribuer des fonctions aux protéines qui sont similaires à une protéine de référence dont on connaît la fonction. Toutefois, cette méthode présente des limites importantes. D'une part, elle ne permet pas de prédire des fonctions pour des protéines qui ne sont similaires à aucune protéine de référence (on dispose de moins de 30% de protéines annotées dans la plupart des datasets). D'autre part, la prédiction est peu précise pour les protéines très similaires à plusieurs protéines de référence ayant des fonctions différentes.

Les récentes avancées en apprentissage de représentation appliquées à la bio-informatique pour la prédiction de fonctions permettent de générer des embeddings (c'est-à-dire des vecteurs denses) pour des séquences protéiques (*cf.* SeqVec [6], ProtBert [4] ou ProteinBert [2]). Ces embeddings permettent également de calculer la similarité entre deux protéines et de prédire des fonctions. Bien que la qualité de la prédiction par similarité d'embeddings dépasse celle obtenue par similarité de séquences, les limites demeurent pour les protéines trop éloignées d'une protéine de référence.

De plus, le transfert d'annotation basé sur un embedding moyen unique par protéine présente des limites fondamentales : cette agrégation efface une grande partie de l'information locale portée par chaque résidu. Associer plusieurs embeddings par protéine — un par résidu — permet au contraire de capturer des signaux locaux faibles révélateurs de similarités fonctionnelles fines, difficilement détectables autrement [9].

1. cf. par exemple <https://docs.mgnify.org/src/docs/mgnify-proteins.html>

Au-delà de l'information de séquence, la structure tridimensionnelle constitue une source complémentaire pour la prédiction de fonction. Les récentes avancées en apprentissage automatique permettent désormais de prédire cette structure avec une grande précision, notamment grâce à AlphaFold [7]. Toutefois, l'information structurale seule ne suffit pas toujours : cette limite apparaît lorsque la structure prédite diffère de celles de protéines de fonction connue, ou lorsque des structures similaires correspondent à des fonctions biologiques distinctes.

Objectifs

La gestion de données biologiques hétérogènes (séquences, embeddings, structures 3D prédites, métadonnées écologiques) pose de grands défis en termes d'intégration multimodale et de scalabilité. Ces défis se posent en particulier lorsqu'il s'agit d'inférer automatiquement la fonction de protéines à partir de telles données. L'objectif général de cette thèse est de concevoir de nouvelles méthodes de prédiction d'annotations fonctionnelles dans des graphes hétérogènes massifs et bruités, en intégrant conjointement différentes mesures de similarité entre nœuds et les informations sémantiques associées à chaque nœud.

Ces méthodes devront satisfaire deux critères complémentaires : améliorer significativement l'état de l'art en termes de précision de prédiction, et passer à l'échelle afin de s'appliquer à de très grands graphes contenant plusieurs milliards de nœuds.

Méthode

Hypothèse de travail

Nous faisons l'hypothèse qu'au sein d'un groupe de protéines similaires par leur alignement séquentiel et/ou leur représentation vectorielle, il existe des corrélations entre leur structure tridimensionnelle et leur fonction. Plus précisément, nous postulons l'existence de motifs structuraux — au sens de configurations géométriques locales de la structure 3D — qui caractérisent des divergences fonctionnelles au sein d'un tel groupe, et qui ne sont pas capturés par les méthodes de transfert d'annotation fondées sur la seule information de séquence.

Plan de travail

La prédiction s'effectuera par analyse collective des attributs multimodaux des protéines (séquence, structure, contexte écologique), plutôt que par simple transfert d'homologie. Pour opérationnaliser cette hypothèse, une variance structurale intra-groupe élevée sera interprétée comme le reflet de motifs structuraux discriminants — et donc comme un marqueur de divergence fonctionnelle — tandis qu'une variance faible caractérisera une fonction conservée. Le plan de travail est le suivant :

Construction des graphes hétérogènes. La première étape consiste à modéliser les protéines sous forme de graphes attribués multi-couches, où chaque nœud représente une protéine et agrège trois modalités de données : sa séquence (via des embeddings issus de modèles de langage protéiques), sa structure 3D (prédite par AlphaFold), et son contexte écologique (métadonnées environnementales). Deux types d'arêtes relient ces nœuds : des arêtes séquentielles encodant les relations d'homologie, et des arêtes structurales encodant la similarité géométrique 3D. Cette représentation multi-couches permettra d'exploiter conjointement les signaux complémentaires portés par chaque modalité lors des étapes d'analyse et d'apprentissage ultérieures.

Analyse structurelle de groupes. Il s'agira d'identifier, au sein de superfamilles protéiques, les ensembles partageant des éléments structuraux communs, et de détecter les motifs structuraux discriminants de la fonction, au-delà de la simple conservation séquentielle. Sur cette base, un

algorithme de propagation de labels sur graphes sera proposé, intégrant la détection de motifs structuraux marquant les transitions fonctionnelles. Cet algorithme permettra notamment de résoudre les cas ambigus, c'est-à-dire les protéines séquentiellement similaires à des protéines de référence multifonctionnelles.

Apprentissage de représentations multimodales. L'objectif est de concevoir une architecture capable d'intégrer conjointement trois modalités de données : l'information séquentielle et structurale, encodées simultanément via ProstT5 [5] — un modèle bilingue pré-entraîné sur séquences et tokens structuraux 3Di, qui présente l'avantage de ne pas nécessiter de prédictions AlphaFold explicites en entrée, ce qui est déterminant pour le passage à l'échelle — les structures 3D complètes prédites par AlphaFold [7], mobilisées lorsqu'elles sont disponibles pour des protéines d'intérêt particulier, et les contextes écologiques (métadonnées environnementales).

L'enjeu central est d'apprendre une *distance fonctionnelle effective*, où la similarité entre représentations reflète la convergence fonctionnelle plutôt que la seule homologie séquentielle. La fusion multimodale sera réalisée via une architecture encodeur-décodeur, optimisée par fine-tuning [3], dont les mécanismes d'attention apprendront à pondérer dynamiquement chaque modalité selon le contexte de prédiction — privilégiant par exemple l'information structurale pour distinguer des protéines séquentiellement similaires mais fonctionnellement divergentes.

Passage à l'échelle. Un objectif transversal, de nature principalement algorithmique, sera de garantir que les méthodes développées dans cette thèse puissent s'appliquer à très grande échelle. Cet objectif sera atteint en combinant de manière cohérente des techniques établies dans un pipeline unifié : pré-filtrage par sketching (MinHash), indexation approximative des embeddings (FAISS, HNSW avec compression IVF-PQ), et clustering distribué par décomposition hiérarchique du graphe [8, 1]. La contribution ne porte pas sur ces composantes individuellement, mais sur leur intégration dans le contexte spécifique de graphes protéiques hétérogènes et multimodaux à l'échelle du milliard de nœuds.

Validation expérimentale. Les méthodes seront validées sur des jeux de référence (séquences de dinoflagellés, projet ANR équipe DIVEDINO) puis appliquées à l'échelle métagénomique (TARA Océans). En partenariat avec l'équipe ABI de l'Institut de Systématique, Évolution, Biodiversité (UMR 7205), une validation expérimentale ciblée sera menée dans le cadre du projet **Func-sea du PEPR ATLASEA** dont l'ABI est partenaire, pour la prédiction de fonctions de protéines marines dont l'annotation échappe aux méthodes classiques.

Résultats attendus

Sur le plan scientifique, cette thèse vise à produire trois types de résultats. Premièrement, une amélioration significative de la précision de la prédiction de fonction par rapport aux méthodes de transfert d'annotation existantes, en particulier pour les protéines similaires à des protéines de référence multifonctionnelles. Deuxièmement, la mise en évidence de motifs structuraux discriminants au sein de groupes de protéines séquentiellement similaires, validant ainsi l'hypothèse centrale de la thèse. Troisièmement, un pipeline scalable et open-source pour l'annotation fonctionnelle à large échelle, applicable à des bases de plusieurs milliards de protéines telles que TARA Oceans.

Ces résultats feront l'objet de publications dans des conférences et journaux internationaux en bases de données et en bio-informatique. La collaboration avec l'équipe ABI du MNHN sera poursuivie et portera sur deux aspects complémentaires : la définition a priori de caractéristiques biologiquement plausibles pour les motifs discriminants fonctionnels que la thèse vise à identifier, et la validation biologique des motifs effectivement découverts, par confrontation avec des experts du domaine.

Références

- [1] Yuhe BAI, Camélia CONSTANTIN et Hubert NAACKE. “Leiden-Fusion Partitioning Method for Effective Distributed Training of Graph Embeddings”. In : *Machine Learning and Knowledge Discovery in Databases. Research Track - European Conference, ECML PKDD*. T. 14947. 2024, p. 366-382.
- [2] Nadav BRANDES et al. “ProteinBERT : a universal deep-learning model of protein sequence and function”. In : *Bioinformatics* 38.8 (2022), p. 2102-2110.
- [3] Andrew DICKSON et Mohammad RK MOFRAD. “Fine-tuning protein embeddings for functional similarity evaluation”. In : *Bioinformatics* 40.8 (2024), btae445.
- [4] Ahmed ELNAGGAR et al. “ProtTrans : toward understanding the language of life through self-supervised learning”. In : *IEEE transactions on pattern analysis and machine intelligence* 44.10 (2021), p. 7112-7127.
- [5] Michael HEINZINGER et al. “Bilingual language model for protein sequence and structure”. In : *NAR Genomics and Bioinformatics* 6.4 (2024), lqae150.
- [6] Michael HEINZINGER et al. “Modeling aspects of the language of life through transfer-learning protein sequences”. In : *BMC bioinformatics* 20.1 (2019), p. 723.
- [7] John JUMPER et al. “Highly accurate protein structure prediction with AlphaFold”. In : *nature* 596.7873 (2021), p. 583-589.
- [8] Hubert NAACKE, Noé CURÉ et Olivier CURÉ. “PCCMerge : a parallel method based on merging partial connected components in large graphs”. In : *IEEE BigData Next-Generation Big Graph Learning and Analytics (Next-Gen-BiGLA)*. Washington D.C., USA, 2024, p. 10.
- [9] Lorenzo PANTOLINI et al. “Embedding-based alignment : combining protein language models with dynamic programming alignment to detect structural similarities in the twilight-zone”. In : *Bioinformatics* 40.1 (2024), btad786.