UNIVERSITÉ DE RENNES 1

Direction de la Recherche et de l'Innovation

Appel à projets - Année 2023

Contrats doctoraux «Ordinaires»

ECOLE DOCTORALE : Matisse **POLE DE RECHERCHE : Mathématiques-Numérique**

Unité de recherche d'accueil

Unité de Recherche : IRISA

Equipe : Dyliss

Encadrement

Directeur de thèse : Olivier DAMERON **Unité de recherche** : IRISA

Employeur/Établissement d'origine : Université de Rennes

HDR : • oui **Nb de thèses en cours** : 1 (10%, soutenance prévue en 2023)

Mail : olivier.dameron@irisa.fr **Tél**. : +33 (0) 2 99 84 74 46

Co-directeur de thèse : François COSTE **Unité de recherche** : IRISA

Employeur/Établissement d'origine : Inria

HDR : • non **Nb de thèses en cours** : 1 (40%, soutenance prévue en 2023)

Mail : francois.coste@inria.fr **Tél.** :+33 (0) 2 99 84 74 91

Sujet

Titre : Apprentissage de modèles avec dépendences explicites entre résidus pour la prédiction fonctionnelle de séquences protéiques

Mots clés : Apprentissage automatique, bioinformatique, protéine, séquence, fonction, interprétabilité « by design », frugalité

Cotutelle internationale (oui/non) : non **Si oui, préciser l'établissement et le pays de rattachement** :

Descriptif (15-30 lignes) :
Contexte socioéconomique et scientifique ; Les hypothèses et questions posées ; Les grandes étapes de la thèse et démarche ; Approches méthodologiques et techniques envisagées

Alors que les projets de séquençage permettent d'avoir accès de façon massive aux séquences des protéines de tout le règne du vivant, le problème reste de déterminer la fonction de ces protéines dans l'organisme. A côté de l'effort d'expérimental qui ne passe pas à cette échelle, la prédiction repose aujourd'hui essentiellement sur des banques de signatures caractéristiques de chaque fonction, permettant d'annoter avec la fonction associée les séquences reconnues par une signature. Bien qu'utiles, ces prédictions restent limitées. Une étude a estimé par exemple que même sur une plante modèle particulièrement bien étudiée (A. Thaliana), 40% des enzymes et transporteurs (deux types de protéines à la fonction primordiale dans le métabolisme) portaient une annotation fonctionnelle crédible. L'annotation d'organismes plus lointains et moins connus tels que ceux issus des grands programmes de métagénomique (TARA Ocean et PEPR ATLASea, French Gut, ...) est évidemment plus difficile.

Les signatures utilisées classiquement sont des modèles probabilistes simples modélisant de façon indépendante la distribution des différents acides aminés à chaque position ainsi que les probabilités d'insertions et de suppression de résidus [1]. Dans le cadre de cette thèse, on s'intéressera à étendre cette approche à la prise en compte de dépendances entre les différents acides aminés d'une séquence pour la caractérisation et la prédiction de fonctions. Cette prise en compte a été très fructueuse pour la prédiction de la structure de protéines, notamment avec l'avènement de méthodes permettant de distinguer les couplages directs de ceux indirects entre acides aminés pour prédire les positions en contact (proches en 3D), et donc la structure, qui ont été à la source du succès d'AlphaFold [2]. Nous pensons qu'elle peut être fructueuse aussi pour la caractérisation de fonctions en permettant la capture par ces dépendances, éventuellement à longue distance dans les séquences, d'éléments structuraux importants des protéines pour la fonction. Nous avons déjà eu des premiers résultats dans ce sens lors de la thèse de H. Talibart [3] (soutenue en mars 2021, encadrée par F. Coste) qui a montré que les couplages directs entre acides aminés modélisés par des modèles de Potts permettaient d'améliorer les alignements de séquences protéiques. Nous sommes aussi en train de montrer dans la thèse de N. Buton (soutenance prévue à l'automne 2022, encadrée par Y. Le Cunff et F. Coste) que l'apprentissage de réseaux de neurones profonds de type Transformer [4,5], qui permettent de mieux prendre en compte les dépendances à longue distance grâce à leur mécanisme d'attention, permet d'améliorer significativement les performances en prédiction de fonctions (macro-f1 augmentant de 41% à 54% sur un benchmark de prédiction de 628 classes enzymatiques)[6]. Le coût calculatoire (et environnemental) de cette dernière

approche est cependant très élevé et, même si nous montrons que l'attention peut fournir des élements d'explicabilité, le fonctionnement de ces gigantesques modèles reste très peu compréhensible.

Le sujet de cette thèse vise un meilleur compromis performances-coût-taille-compréhensibilité par l'apprentissage de modèles explicites de dépendances, à la lumière des avancées en apprentissage profond. Le planning prévoit une progression de modèles simples à plus élaborés : on s'intéressera d'abord à la caractérisation de fonctions par des modèles de couplages directs explicites, dans l'esprit des modèles de Potts, puis on pourra s'inspirer d'idées à l'origine des performances des Transformers (pré-entrainement, mécanisme d'attention, inférence des paramètres par programmation différenciable, ajout de couches profondes,...) pour améliorer ces modèles tout en gardant leurs propriétés de représentation explicite et en limitant les besoins en ressources.

Bibliographie :

1. "Biological Sequence Analysis : Probabilistic Models of Proteins and Nucleic Acids", Durbin R, Eddy S R, Krogh A, Mitchison G J., Cambridge University Press, 1998.
2. "Highly accurate protein structure prediction with AlphaFold" Jumper J., Evans R., Pritzel A. *et al.*. *Nature,* 2021.
3. "Comparaison de protéines homologues avec dépendances entre positions par alignement de modèles de Potts", Talibart H, thèse de l'Université de Rennes 1, 2021.
4. "BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding", Devlin J, Chang MW, Lee K, Toutanova K, 2018
5. "ProtTrans: Towards Cracking the Language of Life's Code Through Self-Supervised Learning", Elnaggar, A., Heinzinger, M., Dallago C *et al.*, B., IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021
6. "Predicting enzymatic function of protein sequences with attention", Buton N, Coste F , Le Cunff Y, soumis à Bionformatics