Interdisciplinary PhD project – Bioinformatics – Evolutionary Biology (INS2I-INEE)

Institutions: UMR 9189 CRIStAL and UMR 8198 Evo-Eco-Paleo (EEP) at CNRS/Lille University;

Project title: Algorithms, bioinformatics and evolution for paleoproteomics
Project: Paleo-Info-Evo (CNRS 80|PRIME project)
Promotorship: Hélène Touzet (DR CNRS, bioinformatics), Céline Poux (Associate Professor, Ulille, evolution)
Email contact: helene.touzet|at|univ-lille.fr, celine.poux|at|univ-lille.fr
Start date and duration: October 2023 for 3 years. The starting date may be

Vacancy description
We are pleased to announce a PhD fellowship for a highly motivated, enthusiastic and independent person with a strong interest in the development of bioinformatics algorithms to improve the analysis of proteomic data in paleontological studies. Background knowledge in all or some of these fields are required: evolutionary biology, phylogenetics, sequence analyses and python programing.

Project description
In recent years, the analysis of ancient biological samples has changed our understanding of the evolution of life on Earth, renewing the approaches previously used in paleontology based on the study of fossils or carbon-14 dating. At the forefront of new molecular techniques is paleogenomics (sequencing of ancient DNA), although DNA degrades relatively quickly. More recently, paleoproteomics via ZooArchaeology by mass spectrometry (ZooMS) offers a possibility to identify morphologically ambiguous or unidentifiable bone fragments from bone assemblages. Identification of bones with ZooMS results from the sequencing of a target protein, such as collagen, which is abundant in bone fragments. The collagen present in the samples is digested and the mass of the peptides obtained by spectrometry gives indirect information on the amino acid sequence of the protein studied. To exploit this data, the community works with *marker* peptides, which serve as a sort of molecular barcode for taxonomic assignment. But the use of these marker peptides suffers from two limitations: it remains manual and it neglects the evolutionary dimension of the data. There is therefore a real need to formalize and automate the methods in order to obtain robust and reproducible assignments, even on a large scale. This raises multiple questions:

- How can the marker peptide approach be generalized towards the combination of marker peptides or consensus marker peptides to take full advantage of the phylogenetic signal contained in the data?
- How to infer marker peptides at different taxonomic levels ?
- How to measure the phylogenetic signal contained in the target protein and its peptides ?
- How to reconstruct ancestral protein sequences from spectra and contemporary sequences to enrich contemporary data sets ?

The methods developed will combine sequence algorithmic approaches and a probabilistic framework using protein sequence evolution models to reconstruct phylogenetic trees and ancestral sequences. The expected results are twofold: to develop a toolbox for data analysis, and to propose a methodological framework for an informed use of marker peptides in ZooMS.

Setting and requirements
The project is funded by the CNRS 80|PRIME initiative and will be developed in an inter-institutional and interdisciplinary collaboration between the UMR CRIStAL and UMR EEP of the CNRS and the University of Lille. Furthermore, this project is realized in close collaboration with Fabrice Bray (MSAP) in charge of the ZooMS platform in Lille, and Patrick Auguste (paleontologist, EEP). Master students that are graduating over the summer are welcome to apply. More information on studying at Lille University can be found on the Lille University webpage: https://www.univ-lille.fr/home/international-student/.

Profile of the candidate

- Master's degree in a relevant field: bioinformatics and/or evolution (sequence analysis, Python programming, phylogenetics).
- Eager to acquire new competences and knowledge in proteomics, evolution and/or bioinformatics depending on the candidates' background.
- Ability to work in an interdisciplinary and collaborative environment (independency, reliability, integrity)
- Ability to write clear scientific reports and disseminate results
- Have good non-academic attributes (e.g., maturity, open-mindedness, respectfulness)

Interested?
To apply for this position, please send, to both of the email addresses indicated above, the following information: a complete CV including grades obtained during the Master program; a letter of motivation that also briefly outlines past research accomplishments and future goals; the name and contact information of a previous project supervisor (bachelor or master thesis). Informal inquiries regarding this vacancy can be sent as well to these two email addresses.

References:
- Age estimates for hominin fossils and the onset of the Upper Palaeolithic at Denisova Cave. Nature, 2019
-Species identification of ancient Lithuanian fish remains using collagen fingerprinting. Journal of Archaeological Science, 2018
-Distinguishing African bovids using Zooarchaeology by Mass Spectrometry (ZooMS): New peptide markers and insights into Iron Age economies in Zambia. Plos One, 2021
-Extinct species identification from late middle Pleistocene and earlier Upper Pleistocene bone fragments and tools not recognizable from their osteomorphological study by an enhanced proteomics protocol. Archeometry, 2022
-compareMS2 2.0: An Improved Software for Comparing Tandem Mass Spectrometry Datasets. Journal of Proteomics, 2023
-Semi-supervised machine learning for automated species identification by collagen peptide mass fingerprinting. BMC Bioinformatics, 2018
-An experimental phylogeny to benchmark ancestral sequence reconstruction. Nature Communications, 2016
-Beyond mass spectrometry, the next step in proteomics. Sciences Advances, 2020

Projet de doctorat interdisciplinaire – Bioinformatique – Biology évolutive (INS2I-INEE)

Institutions : UMR 9189 CRIStAL et UMR 8198 Evo-Eco-Paleo (EEP), CNRS/ Université de Lille

Intitule du projet : Algorithmes bioinformatiques et modèles évolutifs pour la paléoprotéomique.
Projet : Paleo-Info-Evo (CNRS 80|PRIME)
Direction de thèse : Hélène Touzet (DR CNRS en bioinformatique), Céline Poux (Maîtresse de Conférences en évolution, Ulille)
Contact : helene.touzet|at|univ-lille.fr, celine.poux|at|univ-lille.fr
Date de début et durée : 1 Octobre 2023 ; 3 années.

Description du poste
Nous avons le plaisir d'annoncer un contrat doctoral pour une personne motivée, enthousiaste et indépendante, fortement intéressée par le développement d'algorithmes bioinformatiques visant à améliorer l'analyse des données protéomiques pour les études paléontologiques.
Des connaissances de base dans tout ou partie des domaines suivants sont demandées : biologie évolutive, phylogénétique, analyses de séquences et programmation en python.

Description du projet
Ces dernières années, l'analyse d'échantillons biologiques anciens a modifié notre compréhension de l'évolution de la vie sur Terre, renouvelant les approches précédemment utilisées en paléontologie, basées sur l'étude des fossiles ou la datation au carbone 14. Au premier rang des nouvelles techniques moléculaires figure la paléogénomique (séquençage de l'ADN ancien). Plus récemment, la paléoprotéomique via la ZooArchaeology by mass spectrometry (ZooMS) offre la possibilité d'identifier des fragments d'os morphologiquement ambigus ou non identifiables à partir d'assemblages d'os. L'identification des os par ZooMS résulte du séquençage d'une protéine cible, telle que le collagène, qui est abondante dans les fragments d'os. Le collagène présent dans les échantillons est digéré et la masse des peptides obtenus par spectrométrie donne une information indirecte sur la séquence des acides aminés de la protéine étudiée. Pour exploiter ces données, la communauté travaille avec des peptides marqueurs, qui servent en quelque sorte de code-barres moléculaire pour l'assignation taxonomique. Mais l'utilisation de ces peptides marqueurs souffre de deux limitations : elle reste manuelle et elle néglige la dimension évolutive des données. Il y a donc un réel besoin de formaliser et d'automatiser les méthodes afin d'obtenir des assignations robustes et reproductibles, même à grande échelle. Cela soulève de multiples questions :

● Comment généraliser l'approche par peptides marqueurs en allant vers la combinaison de peptides marqueurs ou des peptides marqueurs consensus pour tirer pleinement partie du signal phylogénétique contenu dans les données ?
● Comment inférer automatiquement des peptiques marqueurs à différents niveaux taxonomiques ?
● Comment mesurer le signal phylogénétique contenu dans la protéine cible et ses peptides ?
● Comment reconstruire les séquences protéiques ancestrales à partir des spectres et des séquences contemporaines pour enrichir les jeux de données contemporains ?

Les méthodes développées combineront des approches algorithmiques séquentielles et un cadre probabiliste utilisant des modèles d'évolution des séquences protéiques pour reconstruire les arbres phylogénétiques et les séquences ancestrales. Les résultats attendus sont doubles : développer une boîte à outils pour l'analyse des données, et proposer un cadre méthodologique pour une utilisation éclairée des peptides marqueurs dans ZooMS.

Cadre et exigences

Le projet est financé par l'initiative 80|PRIME du CNRS et sera développé dans le cadre d'une collaboration inter-institutionnelle et interdisciplinaire entre l'UMR CRIStAL et l'UMR EEP du CNRS et l'Université de Lille. De plus, ce projet est réalisé en étroite collaboration avec Fabrice Bray (MSAP) en charge de la plateforme ZooMS à Lille, et Patrick Auguste (paléontologue, EEP). Les étudiants en Master qui obtiendront leur diplôme au cours de l'été sont invités à postuler. Plus d'informations sur les études à l'Université de Lille sont disponibles sur la page web : https://www.univ-lille.fr

Profil du candidat

- Master dans un domaine pertinent pour le projet : bioinformatique et/ou évolution (analyse de séquences, programmation Python, phylogénétique)
- Enthousiasme pour acquérir de nouvelles compétences et connaissances en protéomique, évolution et/ou bioinformatique en fonction du parcours du candidat.
- Capacité à travailler dans un environnement interdisciplinaire et collaboratif (indépendance, fiabilité, intégrité)
- Capacité à rédiger des rapports scientifiques clairs et à diffuser les résultats
- Avoir de bonnes qualités non académiques (maturité, ouverture d'esprit, respect)

Vous êtes intéressé.e ?

Afin de postuler à ce poste, veuillez envoyer, aux deux adresses électroniques indiquées ci-dessus, les informations suivantes : un CV complet les relevés de notes obtenues au cours des deux années de Master ; une lettre de motivation qui indique vos premiers résultats en recherche et vos objectifs professionnels ; le nom et les coordonnées de l'encadrant de votre projet de stage de recherche de Master 2. Toute question concernant ce poste peut être envoyée aux deux adresses électroniques indiquées ci-dessus.

Références bibliographiques:

-Age estimates for hominin fossils and the onset of the Upper Palaeolithic at Denisova Cave. Nature, 2019
-Species identification of ancient Lithuanian fish remains using collagen fingerprinting. Journal of Archaeological Science, 2018
-Distinguishing African bovids using Zooarchaeology by Mass Spectrometry (ZooMS): New peptide markers and insights into Iron Age economies in Zambia. Plos One, 2021
-Extinct species identification from late middle Pleistocene and earlier Upper Pleistocene bone fragments and tools not recognizable from their osteomorphological study by an enhanced proteomics protocol. Archeometry, 2022
-compareMS2 2.0: An Improved Software for Comparing Tandem Mass Spectrometry Datasets. Journal of Proteomics, 2023
-Semi-supervised machine learning for automated species identification by collagen peptide mass fingerprinting. BMC Bioinformatics, 2018
-An experimental phylogeny to benchmark ancestral sequence reconstruction. Nature Communications, 2016
-Beyond mass spectrometry, the next step in proteomics. Sciences Advances, 2020