

Méthodes numériques pour le criblage d'échantillons archéologiques et paléontologiques

Encadrants

- Fabrice BRAY (MSAP, ingénieur de recherche CNRS, HDR)
- Hélène TOUZET (CRISTAL équipe BONSAI, DR CNRS, HDR)
- Julie Jacques (CRISTAL équipe ORKAD, MCF)

Objectifs scientifiques

Les progrès récents de la biologie moléculaire ont ouvert la voie à l'accès à l'information génétique contenue dans les échantillons d'organismes anciens et renouvellent approches d'identification des espèces utilisées en archéologie ou en paléontologie. La paléogénomique est à la pointe de ces techniques et a connu un développement spectaculaire. Cependant, ce domaine reste limité par la fragilité de la molécule d'ADN. En revanche, le protéome constitué des protéines anciennes résiduelles contenues dans le matériau tend à présenter une plus grande stabilité dans le temps, ce qui permet d'identifier des espèces à partir d'échantillons remontant à des centaines de milliers d'années [1]. En outre, il s'agit d'une technique rapide, économique et peu invasive. La paléoprotéomique est ainsi un complément intéressant de la paléogénomique, et progresse actuellement à un rythme rapide. Une des premières réalisations a été le criblage à grande échelle des fossiles de la grotte de Denisova à la recherche d'ossements humains à partir du collagène présent [2]. La méthode a depuis été appliquée à un large éventail d'animaux, de périodes et d'écosystèmes [3,4,5]. Il existe aujourd'hui une tendance de fond d'étendre la paléoprotéomique à l'exploration d'autres matériaux, tels que l'émail, la coquille d'œuf, l'ivoire, les cheveux ou les poils, en utilisant d'autres protéines que le collagène comme marqueurs, élargissant ainsi les possibilités d'application de la technique.

Dans ce contexte, la disponibilité de méthodes fiables et efficaces pour l'analyse automatique des données devient cruciale. Il s'agit de données de spectres de masses, issues soit d'instruments MS ou MS/MS. Les outils bioinformatiques existants sont très parcellaires, et les travaux préliminaires présentés dans [6,7] montrent toute la complexité du problème. Il existe donc un réel besoin de fournir à la communauté scientifique un cadre complet spécifiquement conçu pour relever les défis de l'identification des espèces dans les données protéomiques anciennes. C'est l'objet de ce projet de thèse. Plus précisément, le but est d'introduire un cadre formel générique pour manipuler les spectres de masses et les analyser dans des conditions diversifiées.

Cela comprend les trois grandes questions suivantes:

- Proposer des algorithmes dédiés au pré-traitement des spectres MALDI: nettoyage, calibration, deisotoping, traitement des réplicats, contamination.
- Introduire la notion de "spectre modèle théorique", qui servirait de modèle à un organisme donné, et qui tienne compte des spécificités inhérentes des données (modifications post-traductionnelles, présence de contaminants, isotopes), puis proposer des algorithmes pour la génération de spectres théoriques à partir de données de séquences, complètes ou incomplètes, ou à partir de données de masses. Les spectres théoriques devront dans ce cadre réaliser un équilibre sensibilité/spécificité afin à la fois d'atteindre une résolution taxonomique fine et d'être tolérant aux erreurs expérimentales.
- Proposer des algorithmes d'identification d'espèces à travers la comparaison de spectres expérimentaux et des spectres théoriques. Cette comparaison pourra se faire de plusieurs manières: corrélation, arbre de décision ou d'autres méthodes de classification supervisée

Contexte de la thèse

Ce projet s'inscrit dans le cadre d'une collaboration pluridisciplinaire démarrée en 2020 entre le laboratoire de recherche en sciences informatique CRISTAL (<https://www.cristal.univ-lille.fr/>) et la plateforme de Paléoprotéomique de l'unité MSAP (<https://msap-lab.fr/paleoproteomique-et-zooms>). Cette collaboration a donné lieu à la création du logiciel PAMPA (<https://github.com/touzet/pampa>), présenté à la conférence internationale IZAZ 2024 (<https://izaz2024.sciencesconf.org/?lang=en>). CRISTAL réunit plus spécifiquement deux équipes: l'équipe Bonsai, spécialisée en bioinformatique et l'équipe ORKAD, experte en apprentissage pour l'optimisation combinatoire.

Le projet pourra également bénéficier d'échanges avec d'autres laboratoires au niveau national et international.

Programme de travail

- Année 1: revue bibliographique, prise en main des données, constitution de jeux de données, test de l'existant, prototypage des « spectres modèles théoriques » à partir de séquences connues, algorithme de comparaison (système de score). Assignment de spectres MS et MS/MS.
- Année 2: inférence de spectres modèles théoriques en l'absence de séquences: méthodes et évaluation. Développement logiciel.
- Année 3: Valorisation, collaborations applicatives, rédaction de la thèse

Bibliographie

- [1] Paleoproteomics. C. Warinner, K. Korsow Richter, M.J. Collins. Chemical Reviews (2022), <https://pubs.acs.org/doi/10.1021/acs.chemrev.1c00703>
- [2] Identification of a new hominin bone from Denisova Cave, Siberia using collagen fingerprinting and mitochondrial DNA analysis, S. Brown et al. Scientific Reports (2016), <https://www.nature.com/articles/srep23559>
- [3] 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, F. Bray, S. Flament, G. Abrams, D. Bonjean, C. Rolando, C. Tokarski, P. Auguste. Archeometry (2023), <https://onlinelibrary.wiley.com/doi/10.1111/arcm.12800>
- [4] Robust High-Throughput Proteomics Identification and Deamidation Quantitation of Extinct Species up to Pleistocene with Ultrahigh-Resolution MALDI-FTICR Mass Spectrometry. F. Bray, I. Fabrizi, S. Flament, J.L. Loch, P. Antoine, P. Auguste, C. Rolando. Analytical Chemistry (2023), <https://doi.org/10.1021/acs.analchem.2c03301>.
- [5] Peptide mass fingerprinting of preserved collagen in archaeological fish bones for the identification of flatfish in European waters. K. Dierickx, S. Presslee, R. Hagan, T. Oueslati, J. Harland, J. Hendy, D. Orton, M. Alexander, V. Harvey. R. Soc. open sci. (2022), <http://doi.org/10.1098/rsos.220149>
- [6] SPIN enables high throughput species identification of archaeological bone by proteomics. P. Rüther et al. Nature Communications (2022), <https://www.nature.com/articles/s41467-022-30097-x>
- [7] How to get your goat: automated identification of species from MALDI-ToF spectra S. Hickinbotham et al. Bioinformatics (2020) 10.1093/bioinformatics/btaa181