Stage Master M2 : Exploitation de l'analyse topologique de données (TDA) pour l'étude de micro-organismes d'intérêt alimentaire par spectrométrie OPTIR (IR et Raman)

P.-Y. Louis, J.-M. Perrier-Cornet, P. Winckler

Type de contrat : Stage Master 2
Durée : 6 mois
Lieu : Dijon, France
Date de début : Mars 2026 (possibilité de décalage d'un mois)

1 Environnement de recherche : UMR PAM et l'Équipe PMB

Vous intégrerez l'**UMR PAM** (Procédés Alimentaires et Microbiologiques) au sein de l'Équipe Procédés Microbiologiques et Biotechnologiques (**PMB**). L'équipe PMB est pluridisciplinaire (microbiologistes, modélisateurs, mathématiciens, *etc.*) et ses travaux se concentrent sur l'étude des **réponses cellulaires aux perturbations physiques, chimiques et biologiques**. La thématique scientifique de l'UMR PAM concerne l'amélioration de la stabilité et de la conservation des micro-organismes d'intérêt alimentaire. Vous travaillerez en lien direct avec le **Plateau d'Imagerie Spectroscopique (PIMS/DImaCell)**. La Plateforme **DImaCell** (Dispositif d'Imagerie Cellulaire Bourgogne-Franche-Comté) est spécialisée en imagerie cellulaire fonctionnelle appliquée à la Biologie et fournit des technologies de pointe, notamment la microscopie Infra-Rouge et Raman.

2 Contexte du projet : repousser les limites de l'analyse de spectres

L'étude des micro-organismes (tels que bactéries, levures) est cruciale dans le secteur alimentaire. La spectrométrie vibrationnelle **OPTIR** combine les spectres Infra-Rouge (IR) et Raman dans une même étape d'acquisition, permettant de déterminer la composition moléculaire des échantillons.
Les données spectrales sont de haute dimension et peuvent être affectées par un bruit élevé ou des décalages, rendant l'analyse par des outils chimiométriques conventionnels (comme l'Analyse en Composantes Principales ou le Clustering Hiérarchique) souvent inefficace pour extraire des structures pertinentes.

Ce stage vise à exploiter l'**Analyse Topologique de Données (TDA)**, un champ émergent qui utilise la topologie algébrique pour fournir une étude rigoureuse et quantitative de la "forme" et de la structure des données massives. La TDA a déjà démontré son potentiel dans l'analyse de données spectroscopiques de bactéries. Ce projet fait suite aux recherches engagées dans le cadre du projet ANER BFC DATA-MODSTO4FOOD.

3 Missions et objectifs du stage

L'objectif principal est d'appliquer les outils de la TDA aux données spectrales complexes (IR et Raman) des micro-organismes pour en extraire des signatures robustes et interprétables.

3.1. Exploration topologique et détection de sous-populations (Clustering)

- Application de l'**algorithme Mapper** (outil d'exploration de données basé sur les concepts topologiques) pour visualiser la structure globale des ensembles de données spectrales combinées (IR/Raman).
- Identification de sous-groupes ou sous-populations de micro-organismes (souches, états physiologiques, etc.) que les méthodes classiques pourraient masquer ou ne pas détecter.

3.2. Analyse de robustesse (Homologie persistante)

- Utilisation de l'**Homologie persistante (PH)** pour générer des signatures topologiques multi-échelles, appelées diagrammes de persistance.
- Évaluation de l'invariance des signatures topologiques malgré les variations expérimentales courantes des données spectroscopiques OPTIR/Raman (bruit, décalages de longueur d'onde, résolution variable).

3.3. Ingénierie de caractéristiques (*Features*) pour l'apprentissage automatique

- Transformer les diagrammes de persistance en représentations vectorielles exploitables par des modèles d'apprentissage supervisé (classification des espèces/souches).
- Explorer et comparer différentes distances entre diagrammes de persistance, telles que la **distance de Wasserstein** et la **distance Bottleneck**.

4 Profil recherché et compétences

Nous recherchons un étudiant (H/F) de niveau **Bac+5** (Master M2) issu d'une formation spécialisée en statistique, mathématique appliquée et Data Science, ou en éventuellement informatique.

Compétences requises :

- **Compétences mathématiques :** connaissance des concepts de la TDA, incluant l'Homologie Persistante et, idéalement, l'algorithme Mapper.
- **Compétences informatiques :** Bonnes connaissances de la programmation, notamment en `Python` (bibliothèques de TDA comme `GUDHI`, bibliothèques de Machine Learning et idéalement de chimiométrie).
- **Intérêt interdisciplinaire :** Capacité à travailler à l'interface entre science des données et analyse spectroscopique/biologique.

5 Encadrement et valorisation

Encadrement : Le stage sera encadré par P.-Y. Louis (Professeur en mathématiques/statistiques), J.-M. Perrier-Cornet (Professeur en génie des procédés) et P. Winckler (Ingénieur de recherche en Optique et nanotechnologies), responsables de la plateforme d'imagerie cellulaire.
Valorisation : Ce stage offre la possibilité de contribuer à une **publication scientifique** ou à une présentation dans une conférence.

6 Pour postuler

Envoyez votre CV et lettre de motivation aux encadrants du stage.

P.-Y. Louis (PR math/stat) `pierre-yves.louis@institut-agro.fr`

J.-M. Perrier-Cornet (PR génie des procédés) `jean-marie.perrier-cornet@institut-agro.fr`

P. Winckler (IR Optique et nanotechnologies) `pascale.wincklet@institut-agro.fr`

7 Contexte scientifique détaillé et bibliographie

L'étude des micro-organismes (tels que bactéries, levures) est cruciale dans le secteur alimentaire. La spectrométrie vibrationnelle (Infra-Rouge et Raman) est une technique non destructive utilisée pour déterminer la composition moléculaire des échantillons, permettant d'obtenir des informations fondamentales sur la composition chimique des molécules présentes. L'instrumentation OPTIR combine les spectres Infra-Rouge (IR) et Raman dans une même étape d'acquisition, ce qui est d'intérêt pour l'analyse statistique. Les données spectrales issues de ces systèmes combinés sont de haute dimension et peuvent être affectées par un bruit élevé, des décalages spectraux, et des variations de résolution. Ces défis rendent l'analyse par des outils chimiométriques conventionnels (comme l'Analyse en Composantes Principales ou le Clustering Hiérarchique) [9] souvent inefficace pour extraire des structures pertinentes ou des sous-populations cachées.
L'Analyse Topologique de Données (TDA) est un champ émergent [1] qui combine la topologie algébrique et d'autres outils de mathématique pure pour fournir une étude rigoureuse et quantitative de la "forme" et de la structure des données, ce qui est particulièrement utile dans le cadre de données de grande dimension [2, 4]. Les données spectroscopiques sont des données de type fonctionnelles. La TDA est considérée comme une méthode prometteuse pour l'exploration de données massives en chimie analytique et en biologie [6, 8, 5]. Elle a déjà démontré son potentiel dans l'analyse de données spectroscopiques de bactéries (analyse Raman de bactéries uniques, par exemple *Staphylococcus epidermidis*, *Pseudomonas fluorescens* et *Escherichia coli*) [7, 3].

Références

[1] Erik J. Amézquita, Michelle Y. Quigley, Tim Ophelders, Elizabeth Munch, and Daniel H. Chitwood. The shape of things to come : Topological data analysis and biology, from molecules to organisms. *Developmental Dynamics*, 249(7) :816–833, July 2020.

[2] Gunnar Carlsson. Topological methods for data modelling. *Nature Reviews Physics*, 2(12) :697–708, November 2020.

[3] William K. Chang, David VanInsberghe, and Libusha Kelly. Topological analysis reveals state transitions in human gut and marine bacterial communities. *npj Biofilms and Microbiomes*, 6(1) :41, October 2020.

[4] Frédéric Chazal and Bertrand Michel. An Introduction to Topological Data Analysis : Fundamental and Practical Aspects for Data Scientists. *Frontiers in Artificial Intelligence*, 4 :667963, September 2021.

[5] Amish Mishra and Francis Motta. A Pipeline for Data-Driven Learning of Topological Features with Applications to Protein Stability Prediction, August 2024. arXiv :2408.04847 [stat].

[6] Gabriell Máté, Andreas Hofmann, Nicolas Wenzel, and Dieter W. Heermann. A topological similarity measure for proteins. *Biochimica et Biophysica Acta (BBA) - Biomembranes*, 1838(4) :1180–1190, April 2014.

[7] Marc Offroy and Ludovic Duponchel. Topological data analysis : A promising big data exploration tool in biology, analytical chemistry and physical chemistry. *Analytica Chimica Acta*, 910 :1–11, March 2016.

[8] Alexander D. Smith, Paweł Dłotko, and Victor M. Zavala. Topological data analysis : Concepts, computation, and applications in chemical engineering. *Computers & Chemical Engineering*, 146 :107202, March 2021.

[9] Ron Wehrens. *Chemometrics with R : Multivariate Data Analysis in the Natural Sciences and Life Sciences*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.