

M2 Internship – From Bioinformatic Analysis to User Experience: Full-Stack Improvement of a Pipeline for Microbial Identification by MALDI-TOF

SFBI form fields

- **Type:** M2 internship
- **Duration:** 6 months
- **Level:** Master's degree (M2)
- **Location:** INRAE, MaIAGE research unit, StatInfOmics team – Jouy-en-Josas (78), France
- **Stipend:** internship allowance in accordance with current regulations
- **Start date:** January 2027 (flexible)
- **Application deadline:** 30 November 2026
- **Keywords:** MALDI-TOF MS, microbial identification, bacterial genomics, in silico proteomics, GTDB, assignment score, interactive visualization, Python

Description

Background

MALDI-TOF mass spectrometry has become the reference method for routine identification of microorganisms, in clinical microbiology as well as in food and environmental microbiology. A MALDI-TOF spectrum is dominated by abundant proteins of 2 to 20 kDa, mostly ribosomal. A strain is identified by comparing its spectrum against a library of reference spectra. These libraries are largely proprietary and only cover taxa that have already been isolated and cultured, which limits the identification of understudied microorganisms.

The European project MALDIBANK (Horizon Europe, 2025–2029, coordinated by MIRRI-ERIC) brings together four European research infrastructures and about twenty partners in 11 countries. It aims to build an open library of MALDI-TOF spectra covering bacteria, yeasts, filamentous fungi and microalgae (target: 100,000 spectra from at least 20,000 strains), together with the associated identification services. INRAE contributes through its microbial resource collections and its bioinformatics expertise.

An alternative and complementary approach is to predict the spectrum from the genome. GPMsDB (Sekiguchi et al., 2023) computes the theoretical masses of proteins predicted in about 193,000 public bacterial and archaeal genomes, linked to the GTDB taxonomy (release 95), and then matches them against measured peak lists. More than 90% of the tested spectra are correctly identified, at the species or even subspecies level. Two open-source Python tools implement this approach: GPMsDB-dbtch, which builds the database (gene prediction with Prodigal, annotation with HMMER, mass computation), and

GPMsDB-tk, which performs the identification. Three issues currently limit this approach: a genome database dating from 2020, a simplistic prediction of protein masses, and an assignment score that could be improved.

Project

The internship is part of the StatInfOmics team's work on integrating genome-based identification into the MALDIBANK services. It focuses on extending GPMsDB-dbt and GPMsDB-tk along three lines: updating the reference database, improving the reliability of assignments, and making results interpretable by microbiologists. The work will rely on spectrum datasets of known identity: the benchmark datasets published with GPMsDB and the spectra collected within MALDIBANK.

Objectives

- Provide MALDIBANK users with an effective and user-friendly spectrum identification tool.
- For the intern: progressively gain in-depth knowledge of the tool, how it works and the results it produces, in order to design and develop new features, both on the backend (analysis workflow) and on the frontend (UI/UX).

Tasks

Getting started

- Update the tool's database to the latest GTDB version. Release R232 (April 2026) contains 901,341 genomes and 199,923 species, compared with about 32,000 species in the current database. The build pipeline will be adapted to high-performance computing.
- Assess the impact of this update on the benchmark datasets: accuracy at the genus, species and strain levels, and frequency of ambiguous or incorrect identifications. A denser database puts more closely related genomes in competition, which directly shapes the improvement phase.

Improvement

- Improve the annotation of the proteins used as markers (ribosomal proteins and other biomarkers): translation initiation sites, post-translational modifications, expected masses.
- Improve the score used to assign spectra to a strain, species or genus: weighting peaks by their discriminating power, accounting for mass measurement error, and providing a confidence level at each taxonomic rank.

UI/UX design *[Optional, depending on the progress of the previous tasks]*

- Design and implement a contextual interface for visualizing results: matched peaks and corresponding proteins, position of candidates in the taxonomy, assignment confidence.
- Design and integrate additional features: mapping between GTDB and NCBI taxonomies, name history across GTDB releases, etc.

Skills

Required

- M2 in bioinformatics, or in computer science with a strong biological component.
- Good command of Python and the Linux environment (bash); experience with Git.
- Knowledge of microbial genomics (annotation, taxonomy).
- Rigor in performance evaluation (metrics, test datasets) and in code documentation.

Desirable

- High-performance computing (SGE / SLURM) and workflow managers (Snakemake, Nextflow).
- Development of web interfaces or interactive visualizations (Dash/Plotly, Streamlit, D3.js, JavaScript).
- Basic knowledge of mass spectrometry or proteomics.
- Scientific English, as the project takes place within an international consortium.

Autonomy, curiosity and an interest in working with microbiologists and mass spectrometrists will be assets.

Work environment and supervision

The internship will take place in the StatInfOmics team of the MaAGE unit (INRAE, Université Paris-Saclay) in Jouy-en-Josas. The team develops statistical and bioinformatic methods for the analysis of omics data, with a focus on microbial ecology. The intern will have access to the computing resources of the Migale bioinformatics platform.

The internship will involve regular interactions with international users and collaborators within the European MALDIBANK consortium, including follow-up meetings and technical discussions in English.

Supervisors: Sylvain Marthey, Michel-Yves Mistou, Pierre Nicolas (MaAGE, StatInfOmics)].

How to apply

Send a CV, a cover letter and your M1 transcripts to sylvain.marthey@inrae.fr and michel.mistou@inrae.fr, with "M2 Internship GPMsDB" in the subject line. A link to a code repository (GitHub, GitLab) is welcome.

References

- Sekiguchi Y. et al. (2023). A large-scale genomically predicted protein mass database enables rapid and broad-spectrum identification of bacterial and archaeal isolates by mass spectrometry. *Genome Biology* 24:257. <https://doi.org/10.1186/s13059-023-03096-4>
- Parks D.H. et al. (2022). GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Research* 50:D785–D794. <https://doi.org/10.1093/nar/gkab776>
- GPMsDB-tk: <https://github.com/ysekig/GPMsDB-tk> – GPMsDB-dbt: <https://github.com/ysekig/GPMsDB-dbt>