Annual Report Year 1
Anticipating the Chemistry of Life (ACOL)

Marco Visani[1,2], Madleina Caduff[1,2], Daniel Wegmann[1,2], and Pierre-Marie Allard[1,2,3]

[1]Department of Biology, University of Fribourg, 1700 Fribourg, Switzerland
[2]Swiss Institute of Bioinformatics, 1700 Fribourg, Switzerland
[3]Corresponding author, pierre-marie.allard@unifr.ch

November 5, 2025

Contents

1 Introduction

Public databases compiling genes (e.g., GenBank) and proteins (e.g., UniProt) have long been essential resources for the life sciences. Until recently, no equivalent resource existed for small molecules and their producing organisms. The LOTUS[1] project was designed to address this issue, and forms currently the most comprehensive collection of natural products (NP) occurrences. It however gathers compounds physically isolated from species. While this ensures high confidence, the process of extraction and isolation is too slow to realistically document the chemical diversity of life. To address this issue, we explore two directions: experimental and

computational. On one side, the Earth Metabolome Initiative will establish an open, linked knowledge base documenting the metabolome of all living organisms. In parallel, we develop computational strategies to anticipate the chemical constitution of species based on prior knowledge.
My PhD consists in developing a probabilistic model that integrates LOTUS with large-scale mass spectrometry datasets to predict species metabolomes and estimate the likelihood of encountering NPs across the tree of life. After training and validation, I expect the model to predict the likelihood of encountering a given molecule in a given species, even if this compound has never been observed in this species before. I hope that this will help prioritize experimental efforts toward the most promising candidates and that this PhD will contribute to a better understanding of the distribution of specialized metabolites across the tree of life.

2 Markov Random Field

2.1 WP 1

Main project of my PhD.
We assume that the pattern of presence and absence is modulated by similarities within each dimension. For instance, closely related species may share a similar set of NPs and NPs related in their synthesis may share a similar distribution across species. To model such similarities, we adopt a Markov Random Field (MRF) approach.
The first months were dedicated to my C++ training. I started with some simple implementations for the model. We started with some bitwise operations, class creation, familiarization with the tools developed in Daniel Wegmann's lab.
for the different storage vectors. Given the huge amount of data, we store in a single uint64 the state of the node (1 or 0), its coordinate, and the number of times it was a 1. This allowed me to understand some bitwise operations, how to create classes and the use of constructors.
A couple of adjustments were done compared to the model in the grant.

$$\mathbf{\Lambda}_c = \begin{pmatrix} -\alpha_c & \alpha_c \\ 1-\alpha_c & \alpha_c - 1 \end{pmatrix} \tag{1}$$

We changed from μ_0 and μ_1 to a single parameter α which is the change rate to go from 0 to 1.

$$\boldsymbol{P}(n) = \exp(\mathbf{\Lambda}_c \nu_c b(n)). \tag{2}$$

We now have an additional parameter ν_c which is the rate parameter of the clique. With the same prior for all the cliques.
Root state probabilities are given by the stationary distribution of the Markov chain:

$$\boldsymbol{P}_\infty = (1-\alpha_c, \alpha_c)\,. \tag{3}$$

$$\mathbb{P}(L_{ms}|\boldsymbol{x}(m,s), R_{ms}) = \begin{cases} \epsilon & \text{if } \boldsymbol{x}(m,s) = 0, L_{ms} = 1, \\ 1-\epsilon & \text{if } \boldsymbol{x}(m,s) = 0, L_{ms} = 0, \\ R_{ms} & \text{if } \boldsymbol{x}(m,s) = 1, L_{ms} = 1, \\ 1-R_{ms} & \text{if } \boldsymbol{x}(m,s) = 1, L_{ms} = 0. \end{cases} \tag{4}$$

Added an error rate ϵ in case a molecule was wrongly added to LOTUS.

$$R_{ms} = (1 - e^{-\gamma_0 P_m})(1 - e^{-\gamma_1 Q_s}) \tag{5}$$

Changed equation for research effort. Previously if either of the species or the molecules had a lot of papers, then the research effort would directly go to 1. Now with this change, a lot of papers in both molecules *and* species is required for having a high probability.

2.2 WP 2

2.2.1 Context

Implementation of mass spec data still in discussion but we are planning to move from ViMMS [2] to just using outputs of annotation tools. We might use a combination of Sirius, Metfrag, CFM-ID and maybe weight these outputs.

2.3 Fake extract

We would like to model the *loss* of molecules during the extraction process from Nature to the data analysis.
Plate 3 was performed by a bachelor student, but results are weird, so we'll pass everything again. Plate 1 is done.

3 LOTUS Expanded

3.0.1 Context

Along with the PhD, there are some side projects. One of which is the in silico expansion of the molecules in LOTUS thanks to some reaction rules.

3.0.2 Need

This allows to TODO

3.0.3 Task

It turns out me generate some molecules that are true natural products so we would like to see if we can create this synergy between: in silco generation of a compound, create its in silico MS2 spectrum, try to detect them in real data, if we do, then expanded that molecule again.

3.0.4 Object

In silico chemical expansion of molecules in LOTUS, producing $3 \cdot 10^6$ molecules. Website is being build. We want to add the species to the visualization. Still need to figure out how this can be broadly used.

4 EMI-Monorepo

4.0.1 Context

EMI-Monorepo is our rust coded monolith were we are building all the tools for the Earth Metabolome Initiative. Our post doc Luca is leading the software development of the project. We are building everything in Rust from the backend to the frontend, to the data analysis part.

4.0.2 Need

Given the difficulty of the development of the portal, when Luca feels like a project could be a good training for me to learn Rust, he asks me to do it.

4.0.3 Task

4.0.4 Object

Helped Luca on the Monorepo for a couple of crates. Lost a bit the track of what he is doing.

5 Perspectives

I'd like to continue working with Luca in Rust as it is a skill that I want to master by the end of the PhD.

References

[1] A. Rutz et al. "The LOTUS Initiative for Open Knowledge Management in Natural Products Research". In: *eLife* 11 (26, 2022). DOI: 10.7554/eLife.70780.

[2] J. Wandy et al. "ViMMS 2.0: A Framework to Develop, Test and Optimise fragmentation Strategies in LC-MS Metabolomics". In: *Journal of Open Source Software* 7.71 (30, 2022). DOI: 10.21105/joss.03990.