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

Marco Visani[1,2]

[1]Department of Biology, University of Fribourg, 1700 Fribourg, Switzerland
[2]Swiss Institute of Bioinformatics, 1700 Fribourg, Switzerland

November 6, 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 described 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 (bitwise operations, class creation, etc.) and I familiarized myself with the tools developed in Daniel Wegmann's lab.

After having a first version of the model, we performed some simulations to check if the model could infer from data correctly. The simulations allowed us to find some errors in the inference, and make some adjustments to the model. We changed the following equations.

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

The rate matrix as now a single parameter α that models the change rate. This changes the transition probabilities between parent node $p(n)$ and n to :

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

Calculating the root state probabilities then gives :

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

Furthermore, we added an error rate parameter to the probability of LOTUS given the Markov Field:

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

Finally, the probability of discovery R_{ms} was changed to the following:

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

According to the grant, the first working package has been achieved in the correct time frame. We have a first working implementation of the model.

2.2 WP 2

The second working package consists in modelling and integrating mass spectrometry data into the model. We originally started with
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.
As stated in the grant, we want to model the classes of analytes lost at the stages of extraction, liquid chromatography and mass spectrometry fragmentation. To do so, we have a bought a library of 550 pure natural products that we are going to mix together to simulate a biological extract. This will allow us to estimate the loss of molecules between the Nature and the mass spectrometry analysis. Ideally we would like to integrate this loss into the model.
Additionally we want to pass each pure compound in the mass spec. So far 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

One of my side projects is to perform the in silico expansion the molecules present in LOTUS. This allows to create molecules that are not present in LOTUS but that could be potential existing natural products. We then want to fragment in silico these molecules. We hope that this would allow to increase the number of annotations we are able to make on real data. We hope that we will find some of the molecules we have generated in silico in brand-new research publication such as in the Journal of Natural Products. So far, starting from $100'000$ compounds in LOTUS we generated $3 \cdot 10^6$ in silico compounds. As a first proof that the molecules generated are not completely random, there are about $17'000$ of the compounds that are found in COCONUT which is another natural products databases that doesn't report the biological occurrence of the structure.
Using the in silico molecules for MS2 annotation using the molecules from LOTUS would also allow to a certain taxonomical information aboout this in silco molecule.

3.0.2 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.

3.0.3 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.