UNIVERSITY OF FRIBOURG

MASTER THESIS

Anticipating the chemical compositions of organisms across the tree of life.

Author:
Marco VISANI

Supervisors:
Prof. Daniel WEGMANN
Dr. Pierre-Marie ALLARD

A thesis submitted in fulfillment of the requirements
for the degree of Master of Science in Bioinformatics and Computational Biology

in the

Wegmann Group & COMMONS Lab
Department Biology

March 13, 2023

Declaration of Authorship

I, Marco VISANI, declare that this thesis titled, "Anticipating the chemical compositions of organisms across the tree of life." and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date:

"Thanks to my solid academic training, today I can write hundreds of words on virtually any topic without possessing a shred of information, which is how I got a good job in journalism."

Dave Barry

UNIVERSITY OF FRIBOURG

Abstract

Faculty of Science and Medicine
Department Biology

Master of Science in Bioinformatics and Computational Biology

Anticipating the chemical compositions of organisms across the tree of life.

by Marco VISANI

The Thesis Abstract is written here (and usually kept to just this page). The page is kept centered vertically so can expand into the blank space above the title too...

Acknowledgements

The acknowledgments and the people to thank go here, don't forget to include your project advisor...

Contents

List of Figures

List of Tables

List of Abbreviations

LAH	**L**ist **A**bbreviations **H**ere
WSF	**W**hat (it) **S**tands **F**or

List of Symbols

a	distance	m
P	power	W ($\mathrm{J\,s^{-1}}$)
ω	angular frequency	rad

1 Introduction

1.1 Background

Metabolomics is an interdisciplinary field that aims to identify and quantify the small molecules, or metabolites, present in biological systems. By measuring the abundance of metabolites in different organisms, metabolomics researchers can gain insights into the chemical diversity of life on Earth and the biochemical processes underlying it. This information is essential for understanding the interactions between organisms and their environment, as well as for developing new drugs and agricultural products.

Machine learning (ML) algorithms have revolutionized the field of metabolomics by allowing researchers to analyze large and complex data sets. By training ML models on metabolomics data, researchers can identify patterns and relationships between metabolites, and use these patterns to make predictions about the chemical composition of organisms. ML models have been applied to a wide range of metabolomics problems, including disease diagnosis, drug discovery, and ecological studies.

The aim of this thesis is to develop an ML model that can predict the chemical composition of organisms across the tree of life. To achieve this goal, we will use a database of compounds found in different organisms [1], as well as mass spectrometry data that will be used to infer the presence or absence of metabolites in a species. Our model will be trained to identify patterns in the metabolomics data, and use these patterns to make predictions about the presence or absence of metabolites in a given organism.

By developing an accurate and reliable ML model for predicting the chemical composition of organisms, we aim to make important contributions to the fields of metabolomics and chemical ecology. Our study has the potential to facilitate the discovery of novel molecules with therapeutic or industrial applications, as well as to provide new insights into the biochemical diversity of life on Earth.

1.2 Research question and objectives

1.3 Significance of the study

2 Literature Review

2.1 Metabolomics and chemical ecology

2.2 Machine learning in metabolomics

3 Methods

3.1 Model

We seek to infer the presence or absence of M metabolites in S species. Let us define a vector of vectors $\vec{v} = \{\vec{v_1}, \ldots, \vec{v_I}\}$ where $\vec{v_i} = \{v_1, \ldots, v_{J_i}\}$ and J_i is the length of vector $\vec{v_i}$. Since we are interested in at least species and metabolites, we expect $I \geq 2$. Hence for this model we expect $\vec{v} = \{\vec{s}, \vec{m}, \ldots\}$ with $\vec{s} = \{s_1, \ldots, s_S\}$ all species and $\vec{m} = \{m_1, \ldots, m_M\}$ all metabolites. Define a vector $\vec{w}$ where $\vec{w} = \vec{v} \setminus \{\vec{s}, \vec{m}\}$.

Let us further define $\vec{c}$ a vector containing a single element of $\vec{v_i}$ for each $\vec{v_i}$ in $\vec{v}$. Similarly we have $\vec{b} = \vec{c} \setminus \{s, m\}$. We denote by $x_{\vec{c}} = x_{sm\vec{b}}$ whether metabolite m is present ($x_{sm\vec{b}} = 1$) or absent ($x_{sm\vec{b}} = 0$) in species s in properties $\vec{b} \in \vec{w}$.

We denote $\vec{x}_{s\vec{b}} = (x_{s1\vec{b}}, \ldots, x_{sM\vec{b}})$ the vector of molecules present in properties $\vec{b}$ of a specific species s. Finally we define $\vec{x}_s = (\vec{x}_{s1\vec{w}}, \ldots, \vec{x}_{sM\vec{w}})$ the vector of presence/absence of all molecules for each element of each property in $\vec{w}$ for species s.

We assume that related species share a similar set of metabolites and that metabolites related in their synthesis share a similar distribution across species. Let $\mathbb{P}(x_{sm} = 1|\mu_m, \epsilon_{sm\vec{b}})$ be the probability with which metabolite m is present in species s. We then assume that

$$\operatorname{logit} \mathbb{P}(x_{sm} = 1|\mu_m, \epsilon_{sm\vec{b}}) = \mu_m + \epsilon_{sm\vec{b}} \tag{3.1}$$

where μ_m is a metabolite-specific intercept and $\epsilon_{sm\vec{b}}$ is normally distributed with mean 0 and co-variance :

$$\operatorname{cov}(\epsilon_{\vec{c}}, \epsilon_{\vec{c'}}) = \sum_{i=1}^{I} \sum_{d \neq i} \sum_{f=1}^{F} \alpha_{df} \sigma^{(f)}_{c_i c'_i} \tag{3.2}$$

with v the length of vector $\vec{v}$, σ a known measure of covariance and α a positive scalar. $\sigma^{(f)}_{c_i c'_i}$ is defined as a known measure of covariance between property c_i and c'_i when looking at feature f. f being a feature at which the variance is measured. For example, this would allow to discriminate the variance of two species when looked at the "phenotype", or "environment" level or any arbitrary feature one is interested in.

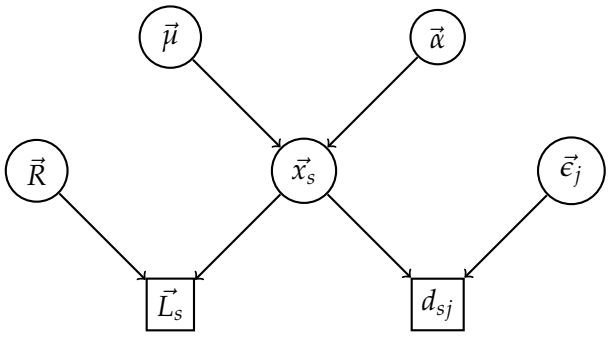

FIGURE 3.1: Potential DAG of the model.

3.2 Origin of data

We have two origins of data, mass spectrometry data and the LOTUS database [1]. We denote d_{sj} the j^{th} mass spectrometry run for species s and $\vec{L_s}$ all molecules assigned to species s present in the LOTUS database. Finally we define R, a function representing the research effort produced for either a species s or a specific molecule m. We also define $\vec{\epsilon_j}$ a vector of error that is specific for each mass-spectrometry run.

A DAG of the model can be seen in Figure 3.1.

3.3 MCMC

To update our parameters using Metropolis-Hastings algorithm, we need to propose a suitable ration h for each parameter to be updated. We know that :

$$P(\vec{\alpha}|\vec{x_s},\vec{\mu}) \propto P(\vec{x_s}|\vec{\mu},\vec{\alpha})P(\vec{\alpha}) \tag{3.3}$$

and,

$$P(\vec{\mu}|\vec{x_s},\vec{\alpha}) \propto P(\vec{x_s}|\vec{\mu},\vec{\alpha})P(\vec{\mu}) \tag{3.4}$$

It is then possible to update each parameter in turn giving :

$$q_\alpha = min\left[1, \frac{P(\vec{x_s}|\vec{\mu},\vec{\alpha\prime})}{P(\vec{x_s}|\vec{\mu},\vec{\alpha})} \cdot \frac{P(\vec{\alpha\prime})}{P(\vec{\alpha})}\right] \quad \text{and,} \quad q_\mu = min\left[1, \frac{P(\vec{x_s}|\vec{\mu\prime},\vec{\alpha})}{P(\vec{x_s}|\vec{\mu},\vec{\alpha})} \cdot \frac{P(\vec{\mu\prime})}{P(\vec{\mu})}\right] \tag{3.5}$$

4 Results

4.1 Presentation of the results of training and testing the ML model

4.2 Analysis of the accuracy and reliability of the model

5 Discussion

5.1 Interpretation of the results in the context of the research question and objectives

5.2 Comparison of the results to previous studies in the field

5.3 Implications of the study for future research and applications in metabolomics

6 Conclusion

6.1 Summary of the main findings and contributions of the study

6.2 Limitations and future directions for research

6.3 Final remarks

A Frequently Asked Questions

A.1 How do I change the colors of links?

The color of links can be changed to your liking using:

`\hypersetup{urlcolor=red}`, or

`\hypersetup{citecolor=green}`, or

`\hypersetup{allcolor=blue}`.

If you want to completely hide the links, you can use:

`\hypersetup{allcolors=.}`, or even better:

`\hypersetup{hidelinks}`.

If you want to have obvious links in the PDF but not the printed text, use:

`\hypersetup{colorlinks=false}`.

References

[1] Adriano Rutz, Maria Sorokina, Jakub Galgonek, Daniel Mietchen, Egon Willighagen, Arnaud Gaudry, James G Graham, Ralf Stephan, Roderic Page, Jiří Vondrášek, Christoph Steinbeck, Guido F Pauli, Jean-Luc Wolfender, Jonathan Bisson, and Pierre-Marie Allard. "The LOTUS Initiative for Open Knowledge Management in Natural Products Research". In: *eLife* 11 (May 2022), e70780. ISSN: 2050-084X. DOI: 10.7554/eLife.70780.

[2] Maha Abdeladhim, Ryan C. Jochim, Melika Ben Ahmed, Elyes Zhioua, Ifhem Chelbi, Saifedine Cherni, Hechmi Louzir, José M. C. Ribeiro, and Jesus G. Valenzuela. "Updating the Salivary Gland Transcriptome of Phlebotomus Papatasi (Tunisian Strain): The Search for Sand Fly-Secreted Immunogenic Proteins for Humans". In: *PLoS ONE* 7.11 (Nov. 2012). Ed. by Luciano A. Moreira, e47347. ISSN: 1932-6203. DOI: 10/f4czvz.

[3] Oliver Baars and David H. Perlman. "Small Molecule LC-MS/MS Fragmentation Data Analysis and Application to Siderophore Identification". In: *Applications from Engineering with MATLAB Concepts*. Ed. by Jan Valdman. InTech, July 2016. ISBN: 978-953-51-2459-7 978-953-51-2460-3. DOI: 10.5772/63018.