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 of Biology

July 18, 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 of 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 Natural Products: Definition and Roles

Natural Products (NPs) are chemical entities biosynthesized by living organisms. Many NPs are metabolites, which can be placed along a specialization gradient from core metabolites that play essential functions and are found in a wide range of organisms to specialized metabolites which are much more restricted across the tree of life.

1.2 The Importance of Natural Products in Therapeutics and Ecosystem Functioning

Specialized metabolites display particular chemical structures and exhibit specific roles and constitute the major part of the current human therapeutic arsenal. Furthermore, their structural characterization and the elucidation of their biological roles are increasingly recognized as fundamental to understanding ecosystems functioning. Their description, however, is not an easy task. Indeed, specialized metabolites are characterized by several levels of complexity. At the scale of a single molecule, their structural complexity explains both their potent biological activities (privileged structures) and their complicated synthetic accessibility. At the organism level, specialized metabolites are found within complex mixtures of extremely diverse chemical classes spanning large dynamic ranges.

1.3 The LOTUS Database and Current Efforts

Some effort have been made to anticipate metabolic networks or occurrences of molecules in selected taxa. A major such resource is the LOTUS database [3] developed and maintained by the COMMONS lab, which currently lists 750'000 thousand occurrences of natural products. However, so far, no model has been proposed to predict their occurrence across the tree of life.

1.4 Project Description and Objectives

The goal of this project is to develop such a model and to train it using large-scale metabolomics and other occurrence data.

2 Literature Review

3 Methodology

3.1 Approach I : Random Markov Field

To model such similarities, we adopt a Markov Random Field approach [1].

Let D denote the total number of dimensions, of which, without loss of generality, the first shall be the metabolite. Each dimension $d = 1, \ldots, D$ consist of a set $\mathcal{E}_d$ of discrete entries (e.g. individual species along the species dimension). We model similarities between the entries of dimension d using a Markov process along a known tree $\mathcal{T}_d$ consisting of $\mathcal{N}_d = \mathcal{E}_d \cup \mathcal{R}_d \cup \mathcal{I}_d$ nodes, of which the entries $\mathcal{E}_d$ are leaves, connected to the set of roots $\mathcal{R}_d$ through a set $\mathcal{I}_d$ of internal nodes; $\mathcal{E}_d \cap \mathcal{R}_d = \varnothing$, $\mathcal{E}_d \cap \mathcal{I}_d = \varnothing$ and $\mathcal{R}_d \cap \mathcal{I}_d = \varnothing$. For every node $n \in \mathcal{N}_d, n \notin \mathcal{R}_d$ that is not a root, let $p(n) \in \mathcal{N}_d$ denote its parent node and $b(n) \geq 0$ the length of the branch connecting it to its parent.

Let $\mathcal{X}$ denote a Markov Random Field of which each variable $x \in \mathcal{X}$ represents a combination of nodes from each of the D dimensions and indicates the presence($x = 1$) or absence ($x = 0$) of metabolite for that combination of nodes. Let $\delta_d(x) \in \mathcal{N}_d$ reflect the node of x in dimension d with $\delta_1(x)$ indicating the metabolite of x, and let $\delta(x) = (\delta_1(x), \ldots, \delta_D(x))$. We only consider two sets of variables: 1) the set $\mathcal{Y}$ of variables representing an entry in each dimension such that for a variable $y \in \mathcal{Y}$, $\delta_d(y) \in \mathcal{E}_d$ for all $d = 1, \ldots, D$, and 2) the set $\mathcal{Z}$ of variables representing leaves in all dimensions except one such that for a variable $z \in \mathcal{Z}$, $\delta_k(z) \in \mathcal{I}_k$ and $\delta_d(z) \in \mathcal{E}_d$ for all $d \neq k$. We then have $\mathcal{X} = \mathcal{Y} \cup \mathcal{Z}$ and $\mathcal{Y} \cap \mathcal{Z} = \varnothing$.

We suppose that the joint density of $\mathcal{X}$ can be factorized over a set of cliques $\mathcal{C}$. Each clique $c \in \mathcal{C}$ consist of a set of variables $x_1, x_2, \ldots \in \mathcal{X}$ that represent the same leaves in all but one dimension k. Specifically, for all $x \in c$, $\delta_d(x) \in \mathcal{E}_d$ for all $d \neq k$ and $\delta_k(x) \in \mathcal{N}_k$, and for all $x_i, x_j \in c$, $\delta_{-k}(x_i) = \delta_{-k}(x_j)$, where $\delta_{-k}(x)$ denotes the vector of nodes of x in all dimensions but k. For such a clique, we will refer to the dimension $\nu(c) = k$ as its *variable* dimension and will denote by $\delta_{-\nu(c)}(c)$ the vector of nodes in the *fixed* dimensions. By definition, $\delta_{-\nu(c)}(c) = \delta_{-\nu(c)}(x)$ for every $x \in c$.

We will further denote by $\mathcal{C}_k \subset \mathcal{C}$ the subset of cliques that share the variable dimension k, i.e. $\nu(c) = k$ for all $c \in \mathcal{C}_k$. Note that each clique is in exactly one subset ($\mathcal{C}_k \cap \mathcal{C}_d = \varnothing$ for all $k \neq d$) and cliques of the same subset do not share any variables ($c_1 \cap c_2 = \varnothing$ for all $c_1, c_2 \in C_k$). However, each variable $x \in \mathcal{Y}$ will be part of exactly one clique from each subset: the clique $c \in \mathcal{C}_k$ for which $\delta_{-k}(c) = \delta_{-k}(x)$. In contrast, each variable $x \in \mathcal{Z}$ will be part of exactly one clique: the clique $c \in \mathcal{C}$ for which $\delta_{-\nu(c)}(c) = \delta_{-\nu(c)}(x)$ and $\delta_{\nu(c)}(x) \in \mathcal{I}_{\nu(c)}$.

The joint density of $\mathcal{X}$ factorizes as

$$\mathbb{P}(\mathcal{X}) = \prod_{d=1}^{D} \prod_{c \in \mathcal{C}_d} \phi(c), \tag{3.1}$$

where we model the clique functions $\phi(c)$ using a Markov model along tree $\mathcal{T}_d$. Let

$$\mathbf{\Lambda}_c = \begin{pmatrix} -\mu_{c1} & \mu_{c1} \\ \mu_{c0} & -\mu_{c0} \end{pmatrix} \tag{3.2}$$

be the rate matrix for changes between states 0 and 1 along the tree. For each node $n \in \mathcal{N}_d, n \notin \mathcal{R}_d$ that is not a root, the transition probabilities between parent node $p(n)$ and n are then given by

$$\boldsymbol{P}(n) = \exp(\mathbf{\Lambda}_c b(n)). \tag{3.3}$$

We assume the root state probabilities are given by the stationary distribution of the Markov chain:

$$\boldsymbol{P}_\infty = \left(\frac{\mu_{c0}}{\mu_{c0} + \mu_{c1}}, \frac{\mu_{c1}}{\mu_{c0} + \mu_{c1}} \right). \tag{3.4}$$

The clique function $\phi(c)$

$$\phi(c) = \prod_{x \in c,} \Big(\mathrm{I}(x \in \mathcal{R}_{\nu(c)}) [\boldsymbol{P}_\infty]_x + \mathrm{I}(x \notin \mathcal{R}_{\nu(c)}) [\boldsymbol{P}(\delta_{\nu(c)}(x))]_{p_c(x),x} \Big) \tag{3.5}$$

where we used the shorthand $x \in \mathcal{R}_{\nu(c)}$ for $\delta_{\nu(c)}(x) \in \mathcal{R}_{\nu(c)}$ to indicate whether the node in the variable dimension of c of x is a root and $p_c(x)$ to identify the variable $z \in c$ for which $\delta_{\nu(c)}(z) = p(\delta_{\nu(c)}(x))$.

3.2 Approach II : Graph Convolution Neural Network (GraphSAGE)

3.3 Approach III : Knowledge Graph Completion

3.4 Data Preprocessing and Model Training

4 Results and Discussion

4.1 Random Markov Field

4.2 Graph Convolution Neural Network (GraphSAGE)

4.3 Knowledge Graph

4.4 Challenges: Data Sparsity and Detection Uncertainty

5 Applications and Implications

6 Conclusion and Future Work

Bibliography

[1] D. Sherrington, S. Kirkpatrick, Solvable Model of a Spin-Glass, Phys. Rev. Lett. 35 (26) (1975) 1792–1796. doi:10.1103/PhysRevLett.35.1792.

[2] W. L. Hamilton, R. Ying, J. Leskovec, Inductive Representation Learning on Large Graphsdoi:10.48550/ARXIV.1706.02216.

[3] A. Rutz, M. Sorokina, J. Galgonek, D. Mietchen, E. Willighagen, A. Gaudry, J. G. Graham, R. Stephan, R. Page, J. Vondrášek, C. Steinbeck, G. F. Pauli, J.-L. Wolfender, J. Bisson, P.-M. Allard, The LOTUS initiative for open knowledge management in natural products research, eLife 11 (2022) e70780. doi:10.7554/eLife.70780.