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 21, 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 [1]. Many NPs are metabolites, which can be positioned along a specialization gradient from core metabolites, fulfilling essential functions and found in a wide range of organisms, to specialized metabolites, much more restricted across the tree of life. Natural Products research is a transdisciplinary field with interests spanning from fundamental structural aspects of naturally occurring molecular entities to their effects on living organisms and extending to the study of chemically mediated interactions within entire ecosystems. We will use *Rutz et al.*'s in [2] definition of Natural Product *as any chemical entity found in a living organism*, *i.e.* a structure-organism pair. This pair holds an additional and fundamental element - a reference to the experimental evidence establishing the linkage between the chemical structure and the biological organism

1.2 The Importance of Natural Products in Therapeutics and Ecosystem Functioning

Unique in their chemical structures and functionalities, specialized metabolites serve as the cornerstone of today's therapeutic applications. Their relevance extends beyond human and veterinary medicine, touching on areas such as agriculture, the food industry, cosmetics, and an array of other sectors. Intrinsically linked to renewable resources, these natural products hold substantial value within the circular economy concept.

The intricacies of their biological roles and structural characterization are crucial to fully grasp the functioning of ecosystems. These complexities influence a myriad of aspects, from the impact on individual organisms to the overall chemically mediated interactions within entire ecosystems. Despite their multifaceted nature, which presents significant challenges at the molecular level - particularly due to their powerful biological activities (privileged structures) and intricate synthetic accessibility the role of these metabolites is indispensable.

Natural products have also contributed to the development of many foundational scientific concepts, including but not limited to stereochemistry, optical activity, regioselectivity, and chirality. Their diverse and complex nature continually

inspires researchers, sparking the creation of innovative tools that can mimic natural processes to control bioregulation mechanisms and address practical challenges [3].

Even though they're complex and challenging to describe, the role of natural products in therapeutic uses and ecosystem functionality cannot be overstated. Current developments aim to unlock this potential more efficiently, emphasizing the ongoing advancements across all sectors associated with natural products.

1.3 The LOTUS Database and Current Efforts

In recent years, efforts have been made to anticipate metabolic networks or occurrences of molecules in selected taxa. A significant resource is the LOTUS database [2] developed and maintained by the COMMONS lab, which currently lists over 750'000 occurrences of natural products. Despite this, no model has been proposed to predict their occurrence across the tree of life. The LOTUS initiative aims to address these shortcomings by consolidating and sharing structure-organism pair information via an open platform, providing transformative potential for natural products research and beyond. This process involves the harmonization, curation, validation, and open dissemination of referenced structure-organism pairs. Furthermore, LOTUS data's embedding into the vast Wikidata knowledge graph facilitates new biological and chemical insights. The contemporary bioinformatic capabilities offered by the LOTUS initiative have the potential to reshape knowledge management, analysis, and interpretation of data in natural products research.

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 Methodology

2.1 Random Markov Field

We seek to infer the presence or absence of metabolites in a group of samples compartmentalized by several discrete dimensions such as e.g. species, tissue or environmental conditions. We assume that the pattern of presence and absence of these metabolites is modulated by similarities within each dimensions. For instance, closely related species may share a similar set of metabolites and metabolites related in their synthesis may share a similar distribution across species. To model such similarities, we adopt a Markov Random Field approach [4,5].

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 $x \in c$.

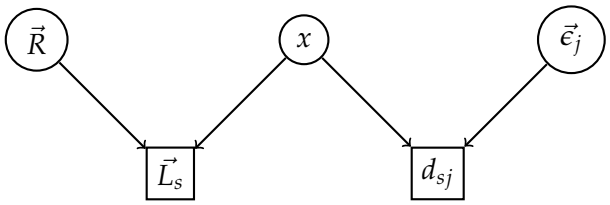

FIGURE 2.1: Potential DAG of the model.

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

$$\mathbb{P}(\mathcal{X}) = \prod_{d=1}^{D} \prod_{c \in \mathcal{C}_d} \phi(c), \tag{2.1}$$

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

$$\mathbf{\Lambda}_c = \begin{pmatrix} -\mu_{c1} & \mu_{c1} \\ \mu_{c0} & -\mu_{c0} \end{pmatrix} \tag{2.2}$$

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

$$\boldsymbol{P}(n) = \exp(\mathbf{\Lambda}_c b(n)). \tag{2.3}$$

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

$$\boldsymbol{P}_\infty = \left(\frac{\mu_{c0}}{\mu_{c0} + \mu_{c1}}, \frac{\mu_{c1}}{\mu_{c0} + \mu_{c1}} \right). \tag{2.4}$$

The clique function $\phi(c)$

$$\phi(c) = \prod_{x \in c,} \Big(\mathrm{I}(x \in \mathcal{R}_{\nu(c)})[\boldsymbol{P}_\infty]_x + \mathrm{I}(x \notin \mathcal{R}_{\nu(c)})[\boldsymbol{P}(\delta_{\nu(c)}(x))]_{p_c(x),x} \Big) \tag{2.5}$$

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 $\delta_{\nu(c)}(z) = p(\delta_{\nu(c)}(x))$.

2.2 Graph Convolution Neural Network (GraphSAGE)

GraphSAGE [6] is an inductive method for learning representations of nodes in graph-structured data. The objective of GraphSAGE is to generate a low-dimensional vector representation for each node in a graph that concisely encapsulates the node's structural and feature information. The central contribution of this method is its capacity for inductive learning, allowing for the generation of embeddings for unseen nodes or even entire unseen graphs, based on the trained model, contrasting it from transductive methods which can only generate embeddings for nodes seen during training.

The architecture of GraphSAGE consists of a series of differentiable aggregator functions which operate in a neighborhood around each node. In the GraphSAGE framework, the learning process occurs in two major steps:

1. **Sampling**: In this phase, a fixed-size neighborhood of each node is sampled. The notion of a 'neighborhood' in this context refers to the set of nodes directly connected to a given node. By limiting the sample size, GraphSAGE ensures that the computational complexity of the subsequent aggregation step remains manageable, even in large-scale graph structures.

2. **Aggregation**: This step involves the aggregation of features from the nodes sampled in the previous step. These features are combined to generate embeddings for the target nodes. The aggregation function may take a number of forms, such as mean, pooling, or Long Short-Term Memory (LSTM) functions, each of which collapses the multi-node feature information into a single vector representation.

By iteratively applying these sampling and aggregation steps across a sequence of layers (corresponding to increasing graph hops), GraphSAGE aggregates information from an expanding neighborhood. The resulting vector embeddings thus encapsulate information about each node's local graph structure and associated feature information.

The flexibility of GraphSAGE allows it to be applied to a wide range of graph-structured data, and its inductive learning capability makes it well-suited to dynamic environments where graph structures may evolve over time or where embeddings for entirely new graphs need to be generated. This represents a significant advantage over traditional transductive graph embedding techniques.

2.3 Knowledge Graph Completion

Knowledge Graphs (KGs) provide a robust technique for the consolidation of diverse data and the modelling of complex interactions, crucial for areas like forecasting the natural products present in various species [7].

In its simplest form, a graph is a data structure that designates items (or "nodes") and the links (or "edges") between them. In the context of our discussion, nodes

might symbolize different species, while edges could denote a variety of relationships such as common ecosystems, shared characteristics, or the potential to yield similar natural products. Depending on the kind of the relationships, the graph can either be directed (representing asymmetric relations) or undirected (symbolizing symmetric relations).

Two primary categories of graphs exist: homogeneous and heterogeneous. A homogeneous graph possesses nodes and edges of the same category. In this scenario, a homogeneous graph might be comprised of species nodes interconnected by edges representing a particular connection, like a mutual ecosystem. Conversely, a heterogeneous graph contains nodes and edges of diverse types. For instance, a heterogeneous graph in this case could encompass nodes representing species, ecosystems, and natural products, linked through various relationships like "inhabits", "generates", or "has common traits with".

A specific kind of graph known as multigraphs permits multiple edges between the same pair of nodes and can also accommodate loops. This is advantageous when there are various types of relations between the same pair of nodes. Multigraphs are predominantly heterogeneous.

A Knowledge Graph (KG) is a distinct type of graph used to encode significant knowledge concerning a specific domain. It is a directed, heterogeneous multigraph with domain-specific semantics for its node and relation types. In this setting, a KG could be employed to encode knowledge about species and their capability to produce natural products. The nodes in the KG, also known as entities, could symbolize different species, ecosystems, or specific natural products. The directed edges, usually represented as triples (head, relation, tail), encapsulate the relationships between these entities.

Knowledge graph embeddings are techniques to convert the discrete entities and relations in a KG into continuous vectors in a high-dimensional space, while preserving the original relationships from the KG [8]. This conversion to a continuous vector space is especially valuable in predicting missing information, like forecasting a species' potential to produce a specific natural product based on its relations with other entities in the KG.

In conclusion, KGs and their corresponding embeddings act as an essential instrument for encoding and analysing multifaceted, relational data. Within the context of predicting natural products in species, they can encapsulate and represent complex relationships between species, ecosystems, and the natural products themselves, potentially contributing to the discovery and comprehension of new natural products.

2.4 Data Preprocessing and Model Training

3 Results and Discussion

3.1 Random Markov Field

3.2 Graph Convolution Neural Network (GraphSAGE)

3.3 Knowledge Graph

3.4 Challenges: Data Sparsity and Detection Uncertainty

4 Applications and Implications

5 Conclusion and Future Work

References

[1] All natural. *Nat Chem Biol*, 3(7):351–351, July 2007. doi:10.1038/nchembio0707-351.

[2] 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. *eLife*, 11:e70780, May 2022. doi:10.7554/eLife.70780.

[3] Pavel B. Drasar and Vladimir A. Khripach. Growing Importance of Natural Products Research. *Molecules*, 25(1):6, December 2019. doi:10.3390/molecules25010006.

[4] David Sherrington and Scott Kirkpatrick. Solvable Model of a Spin-Glass. *Phys. Rev. Lett.*, 35(26):1792–1796, December 1975. doi:10.1103/PhysRevLett.35.1792.

[5] Ross Kindermann and J. Laurie Snell. *Markov Random Fields and Their Applications*. Contemporary Mathematics ; v. 1. American Mathematical Society, Providence, R.I, 1980.

[6] William L. Hamilton, Rex Ying, and Jure Leskovec. Inductive Representation Learning on Large Graphs. 2017. doi:10.48550/ARXIV.1706.02216.

[7] Lisa Ehrlinger and Wolfram Wöß. Towards a Definition of Knowledge Graphs. September 2016.

[8] Quan Wang, Zhendong Mao, Bin Wang, and Li Guo. Knowledge Graph Embedding: A Survey of Approaches and Applications. *IEEE Trans. Knowl. Data Eng.*, 29(12):2724–2743, December 2017. doi:10.1109/TKDE.2017.2754499.