UNIVERSITÉ DE FRIBOURG
UNIVERSITÄT FREIBURG

UNIVERSITY OF FRIBOURG

MANUSCRIPT

ACOL : Random Markov Field

Authors:
Madleina CADUFF
Marco VISANI

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

November 3, 2025

Contents

List of Abbreviations

DAG	**D**irected **A**cyclic **G**raph
GNN(s)	**G**raph **N**eural **N**etwork(s)
MS	**M**ass **S**pectrometry
NP(s)	**N**atural **P**roduct(s)
RMF	**R**andom **M**arkov **F**ield

1 Random Markov Field

1.1 Detailed Research Plan

We seek to infer the presence or absence of NPs 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 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.

Let $D \geq 2$ denote the total number of dimensions, of which, without loss of generality, the first shall be the NP and the second the species. 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: the entries $\mathcal{E}_d$ are leaves, connected to the set of roots $\mathcal{R}_d$ through a set of internal nodes $\mathcal{I}_d$. 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 (see Figure 1.1A for an example).

Markov Random Field 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 the NP for that combination of nodes (see Figure 1.1A for an example with two dimensions). Let $\delta_d(x) \in \mathcal{N}_d$ reflect the node of x in dimension d with $\delta_1(x)$ indicating the NP of x, and let $\delta(x) = (\delta_1(x), \ldots, \delta_D(x))$ be the vector of nodes across all dimension D. We only consider two sets of variables: 1) the set $\mathcal{Y}$ of variables representing a leaf 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 $c \in \mathcal{C}$ consists of a set of variables of $\mathcal{X}$ that represent the same leaves in all but one dimension k. Specifically, for all $x, x' \in c$, $\delta_k(x) \in \mathcal{N}_k$, $\delta_{d \neq k}(x) = \delta_d(x') \in \mathcal{E}_d$ (see Figure 1.1A for an illustration). 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

$$\mathbb{P}(\mathcal{X}) = \prod_{d=1}^{D} \prod_{c \in \mathcal{C}_d} \phi(c). \tag{1.1}$$

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

$$\mathbf{\Lambda}_c = \begin{pmatrix} -\alpha_c & \alpha_c \\ 1-\alpha_c & \alpha_c - 1 \end{pmatrix} \tag{1.2}$$

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

$$\boldsymbol{P}(n) = \exp(\mathbf{\Lambda}_c \nu_c b(n)). \tag{1.3}$$

TODO : change $\nu(c)$ for something else as we have now the scaling rate of the in 1.4

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

$$\boldsymbol{P}_\infty = (1-\alpha_c, \alpha_c)\,. \tag{1.4}$$

The clique function $\phi(c)$ is given by

$$\phi(c) = \prod_{x \in c,} \Big(\mathrm{I}(x \in \mathcal{R}_{\nu(c)})[\boldsymbol{P}_\infty]_x + \mathrm{I}(x \notin \mathcal{R}_{\nu(c)})[\boldsymbol{P}(\delta_{\nu(c)}(x))]_{p_c(x),x} \Big) \tag{1.5}$$

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

Inference & Scalability The number of species and NPs in nature are huge. The model proposed above was carefully crafted to scale to these numbers. The total number of hierarchical parameters to infer are rate parameter per clique, one scaling

factor and one branch length per node of each dimension, regardless of the number or size of the other dimensions. We will infer all parameters under a Bayesian scheme using Markov Chain Monte Carlo (MCMC) techniques to obtain posterior probabilities on all $x \in \mathcal{X}$.

Observed data

We consider two types of data to inform about $\mathcal{X}$: i) presence-only reports of specific NPs in specific species as available through the LOTUS database [1] and ii) presence-absence data obtained with mass-spectrometry (LC-MSMS). We outline emission models for both types of data, but stress that these are mere starting points and may be made more elaborate in the future.

LOTUS Let $L_{ms} = 1$ denote a known occurrence of NP m in species s, and let $L_{ms} = 0$ denote that no evidence for such an occurrence has been reported, either because the NP m is truly absent in species s or because of a lack of research effort.

Let $x(m,s)$ denote the variable in $\mathcal{X}$ for NP m and species s, which, in case $\mathcal{X}$ contains additional dimensions, is obtained by collapsing:

$$x(m,s) = \min\left(1, \sum_{x \in \mathcal{X}} \mathrm{I}(\delta_1(x) = m)\,\mathrm{I}(\delta_2(x) = s)x\right).$$

We will model the probability of L_{ms} given $x(m,s)$ as a function of R_{sm}, the probability of discovery of NP m in species s, such that

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

The probability of discovery R_{ms} will be modeled as a function of the known total number of relevant papers published for NP m (P_m) and for species s (Q_s), such that

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

with positives scalars γ and δ that we will infer.

LC-MSMS Ultra High Performance Liquid Chromatography coupled to fragmentation Mass Spectrometry (LC-MSMS) is the analytical workhorse for the molecular characterization of complex biological matrices. Coupled to computational mass spectrometry tools, such untargeted approaches allows to detect and putatively annotate thousands of NPs in a single run. These analysis are fundamental in our

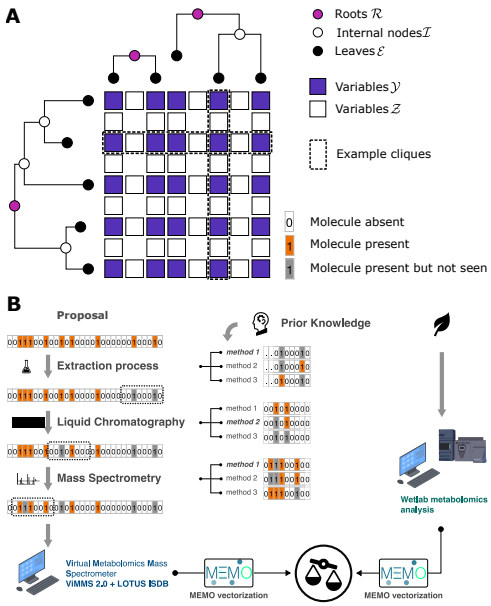

Figure 1.1: Visualizations of core models. A) An illustration of a Markov Random Field for two dimensions (*e.g.* NPs and species). B) Workflow to calculate the probability of the observed mass-spectrometry data given a proposed vector indicating the presence of NPs in a species. Full figure here.

project as they will allow us to complete the currently patchy overview offered by LOTUS. However, the resulting data is complex and requires careful processing to be informative as it is high-dimensional, noisy and incomplete. We here propose a model that builds on previous work regarding the Liquid Chromatography (LC) process [2,3], the establishment of virtual mass spectrometers [4–6] and the integration of LC and MS dimensions in the NP annotation process [7,8], yet is simpler and more streamlined to render it computationally feasible for the large scales considered here.

Suppose $\mathcal{D}_i$ is an LC-MSMS profile obtained from a sample representing a specific vector $\boldsymbol{\xi} = (\xi_{i2}, \ldots, \xi_i D), \xi_{id} \in \mathcal{E}_d$ of leaves in all dimensions except NPs one, such as, for instance, a sample representing a specific tissue of a specific species. We will calculate the probability of the LC-MSMS data $\mathcal{D}_i$ given $\mathcal{X}$ as schematized in Figure 1.1B: Let $\boldsymbol{x}(\boldsymbol{\xi}_i) \subset \mathcal{X}$ denote a slice through $\mathcal{X}$ relevant for $\mathcal{D}_i$, *i.e.* consisting of all variables that represent a leave in the metabolite dimension and the specific leaves $\boldsymbol{\xi}$ in each other dimension such that for all $x, x' \in \boldsymbol{x}(\boldsymbol{\xi}_i)$, $\delta_1(x) \in \mathcal{N}_1$, $\delta_1(x) \neq \delta_1(x')$ and $\delta_{d\neq 1}(x) = \delta_{d\neq 1}(x') = \xi_{id} \in \mathcal{E}_d$. We will develop an NPs-flavored **Vi**rtual **M**etabolomics **M**ass **S**pectrometer (**ViMMS**) building on the original implementation [4]. It will be fed by an in silico spectral database of the last LOTUS contents (ISDB-LOTUS) and informed by prior expert knowledge regarding the classes of analytes lost in the reductionist and stochastic metabolomics approach here formed by *extraction*, *liquid chromatography* and *mass spectrometry fragmentation* stages. The NPs-ViMMS will allow the generation of theoretical metabolomics datasets for any given

input ($\mathcal{D}'_i$), these will be then be compared to experimental results ($\mathcal{D}_i$).

To compare the resulting LC-MSMS profile $\mathcal{D}'_i$ to $\mathcal{D}_i$, we will then take advantage of **MEMO** (MS2 BasEd SaMple VectOrization), a method we recently established for the computationally efficient comparison of large sets of samples based on their LC-MSMS profiles [9]. The first step is to extract fragment ions and neutral losses from each MSMS spectrum binned from the detected features in so-called "documents" using Spec2Vec [10]. Then, for a given sample, all the documents created are aggregated based on word occurrences to form a fingerprint (a MEMO vector). The MEMO strategy exploits the advantages of LC, namely its separation power (thus simplifying the chemical complexity of the sample being analyzed and allowing resolution of isomerisms) while avoiding the disadvantages of RT-based alignment since MEMO vectors contain only mass spectrometry information. Here we'll implement a stochastic comparison of MEMO vectors $\mathcal{M}(\mathcal{D}'_i)$ and $\mathcal{M}(\mathcal{D}_i)$ using a per entry error rate ϵ to be inferred from the data.

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. *eLife*, 11:e70780, May 2022. URL: https://elifesciences.org/articles/70780, doi:10.7554/eLife.70780.

[2] William Heymann, Juliane Glaser, Fabrice Schlegel, Will Johnson, Pablo Rolandi, and Eric Von Lieres. Advanced error modeling and Bayesian uncertainty quantification in mechanistic liquid chromatography modeling. *Journal of Chromatography A*, 1708:464329, October 2023. URL: https://linkinghub.elsevier.com/retrieve/pii/S002196732300554X, doi:10.1016/j.chroma.2023.464329.

[3] Paweł Wiczling, Agnieszka Kamedulska, and Łukasz Kubik. Application of Bayesian Multilevel Modeling in the Quantitative Structure–Retention Relationship Studies of Heterogeneous Compounds. *Analytical Chemistry*, 93(18):6961–6971, May 2021. URL: https://pubs.acs.org/doi/10.1021/acs.analchem.0c05227, doi:10.1021/acs.analchem.0c05227.

[4] Joe Wandy, Vinny Davies, Justin J. J. Van Der Hooft, Stefan Weidt, Rónán Daly, and Simon Rogers. In Silico Optimization of Mass Spectrometry Fragmentation Strategies in Metabolomics. *Metabolites*, 9(10):219, October 2019. URL: https://www.mdpi.com/2218-1989/9/10/219, doi:10.3390/metabo9100219.

[5] Joe Wandy, Vinny Davies, Ross McBride, Stefan Weidt, Simon Rogers, and Rónán Daly. ViMMS 2.0: A framework to develop, test and optimisefragmentation strategies in LC-MS metabolomics. *Journal of Open Source Software*, 7(71):3990, March 2022. URL: https://joss.theoj.org/papers/10.21105/joss.03990, doi:10.21105/joss.03990.

[6] Joe Wandy, Ross McBride, Simon Rogers, Nikolaos Terzis, Stefan Weidt, Justin J. J. Van Der Hooft, Kevin Bryson, Rónán Daly, and Vinny Davies. Simulated-to-real benchmarking of acquisition methods in untargeted metabolomics. *Frontiers in Molecular Biosciences*, 10:1130781, March 2023. URL: https://www.frontiersin.org/articles/10.3389/fmolb.2023.1130781/full, doi:10.3389/fmolb.2023.1130781.

[7] Eric Bach, Simon Rogers, John Williamson, and Juho Rousu. Probabilistic framework for integration of mass spectrum and retention time information in small molecule identification. *Bioinformatics*, 37(12):1724–1731, July 2021. URL: https://academic.oup.com/bioinformatics/article/37/12/1724/6007259, doi:10.1093/bioinformatics/btaa998.

[8] Eric Bach, Emma L. Schymanski, and Juho Rousu. Joint structural annotation of small molecules using liquid chromatography retention order and tandem mass spectrometry data. *Nature Machine Intelligence*, 4(12):1224–1237, December 2022. URL: https://www.nature.com/articles/s42256-022-00577-2, doi:10.1038/s42256-022-00577-2.

[9] Arnaud Gaudry, Florian Huber, Louis-Félix Nothias, Sylvian Cretton, Marcel Kaiser, Jean-Luc Wolfender, and Pierre-Marie Allard. MEMO: Mass Spectrometry-Based Sample Vectorization to Explore Chemodiverse Datasets. *Frontiers in Bioinformatics*, 2:842964, April 2022. URL: https://www.frontiersin.org/articles/10.3389/fbinf.2022.842964/full, doi:10.3389/fbinf.2022.842964.

[10] Florian Huber, Lars Ridder, Stefan Verhoeven, Jurriaan H. Spaaks, Faruk Diblen, Simon Rogers, and Justin J. J. Van Der Hooft. Spec2Vec: Improved mass spectral similarity scoring through learning of structural relationships. *PLOS Computational Biology*, 17(2):e1008724, February 2021. URL: https://dx.plos.org/10.1371/journal.pcbi.1008724, doi:10.1371/journal.pcbi.1008724.