# Inferring the presence of metabolites

March 6, 2023

## 1 Model

### 1.1 Description

We seek to infer the presence or absence of $M$ metabolites in tissue $T$ in $S$ species. We denote by $x_{smt}$ whether metabolite $m = 1, \dots, M$ is present ($x_{smt} = 1$) or absent ($x_{smt} = 0$) in tissue $t = 1, \dots T$ in species $s = 1, \dots, S$. We denote $\vec{x_{st}} = (x_{st1}, \dots, x_{stM})$ the vector of molecules present in a tissue $t$ of a specific species $s$.

Let us further denote $\vec{x_s} = (\vec{x_{s1}}, \dots, \vec{x_{sT}}) = (x_{s11}, \dots, x_{s1M}, x_{s21}, \dots, x_{sTM})$ the vector of presence/absence of all molecules across all tissues 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 $\mathbb{P}(x_{sm} = 1|\mu_m, \epsilon_{sm})$ be the probability with which metabolite $m$ is present in species $s$. We then assume that

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

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

$$\operatorname{cov}(\epsilon_{\vec{c}}, \epsilon_{\vec{c'}}) = \sum_{i=1}^{C} \sum_{d \neq i} \sum_{f=1}^{F} \alpha_{df} \sigma^{(f)}_{c_i c'_i} \tag{2}$$

with $\sigma$ a known measure of covariance and $\alpha$ a scalar. Based on the definition at the top of Section 1.1, we would define $\vec{c} = \{s, m, t\}$ being specific species, molecule and tissue respectively. $\sigma^{(f)}_{c_i c'_i}$ 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.

#### 1.1.1 Updating parameters

Furthermore, we have two origins of data, mass spectrometry data and the Lotus database. 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 1.

With $\vec{\mu} = (\mu_1, \dots, \mu_M)$

### 1.2 LOTUS database

Since LOTUS database [1] has no properties in $\vec{c}$ other than species and molecules, we denote

$$P(\vec{L_S}|\vec{x_s}, \vec{R}) = P(\vec{L_s}|\vec{\xi_s}, \vec{R}),$$

with $\vec{\xi_s} = (\xi_{s1}, \dots, \xi_{sM})$ the vector of presence/absence of all molecules $M$ in species $s$. Furthermore, $\xi_{sm} = min(1, \sum_t^T x_{smt})$ the minimum between 1 and the sum of presence or absence of a molecule across all tissues.

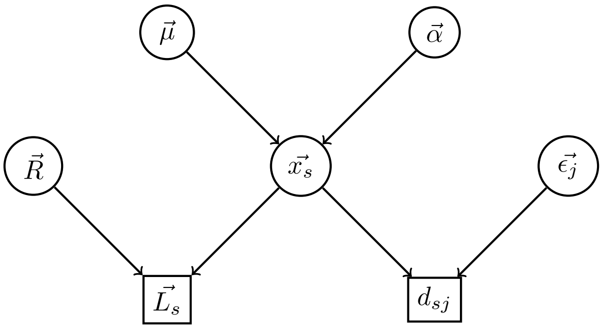

Figure 1: Potential DAG of the model.

### 1.3 MS data

Let $\vec{d_{sj}} = (d_{sj1}, \dots, d_{sjM})$ be the presence-absence vector of each metabolite $m$ obtained with mass-spectrometry run $j = 1, \dots, J_s$ performed on species $s$. Assuming a false-positive and false-negative error rates $\epsilon_{01}$ and $\epsilon_{10}$, respectively, we have

$$\mathbb{P}(\boldsymbol{d}_{sj}|\boldsymbol{x}, \epsilon_{01}, \epsilon_{10}) = \prod_m \left[ x_{sm} \left( \epsilon_{10}^{1-d_{sjm}} (1-\epsilon_{10})^{d_{sjm}} \right) + (1 - x_{sm}) \left( \epsilon_{01}^{d_{sjm}} (1-\epsilon_{01})^{1-d_{sjm}} \right) \right].$$

## 2 Idea and concept

We seek to infer the presence or absence of $M$ metabolites in $S$ species. We denote by $x_{sm}$ whether metabolite $m = 1, \dots, M$ is present ($x_{sm} = 1$) or absent ($x_{sm} = 0$) in species $s = 1, \dots, S$. To infer the full vector $\boldsymbol{x} = (x_{11}, \dots, x_{1M}, \dots, x_{SM})$, 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 $\mathbb{P}(x_{sm} = 1|y_{sm}) = y_{sm}$ be the probability with which metabolite $m$ is present in species $s$. We then assume that

$$\operatorname{logit} y_{sm} = \mu_m + \epsilon_{sm}$$

where $\mu$ is a metabolite-specific intercept and $\epsilon_{sm}$ is normally distributed with mean 0 and co-variance $\operatorname{cov}(\epsilon_{sm}, \epsilon_{s'm'}) = \alpha\sigma_{ss'} + \beta\sigma_{mm'}$ between each combination of species and metabolite. Here, $\sigma_{ss'}$ and $\sigma_{mm'}$ are known measures of covariance between species $s$ and $s'$ and between metabolites $m$ and $m'$, respectively, and $\alpha$ and $\beta$ are positive scalars.

We consider two sets of data informative about $\boldsymbol{x}$: i) Presence-absence data obtained with mass-spectrometry and ii) presence-only reports of specific metabolites in specific specie. Let $\boldsymbol{d}_{sj} = (d_{sj1}, \dots, d_{sjM})$ be the presence-absence vector of each metabolite $m$ obtained with mass-spectrometry run $j = 1, \dots, J_s$ performed on species $s$. Assuming a false-positive and false-negative error rates $\epsilon_{01}$ and $\epsilon_{10}$, respectively, we have

$$\mathbb{P}(\boldsymbol{d}_{sj}|\boldsymbol{x}, \epsilon_{01}, \epsilon_{10}) = \prod_m \left[ x_{sm} \left( \epsilon_{10}^{1-d_{sjm}} (1-\epsilon_{10})^{d_{sjm}} \right) + (1 - x_{sm}) \left( \epsilon_{01}^{d_{sjm}} (1-\epsilon_{01})^{1-d_{sjm}} \right) \right].$$

To model the presence only data, it must be put in relation to the expected research effort. Let $p_{sm}$ denote the known number of presence-only reports for metabolite $m$ in species $s$ and $n_{sm}$ the unknown number of research projects that aimed at discovering metabolite $m$ in species $s$. Assuming a false-positive and false-negative error rates $\pi_{01}$ and $\pi_{10}$, respectively, we have

$$\mathbb{P}(p_{sm}|n_{sm}, \pi_{01}, \pi_{10}) =$$

We would have the covariance matrix such as :

$$\operatorname{cov}(\epsilon_{smt}, \epsilon_{s'm't'}) = \alpha\sigma^P_{ss'} + \beta\sigma^M_{mm'} + \gamma\sigma^E_{ss'} + \dots \tag{3}$$

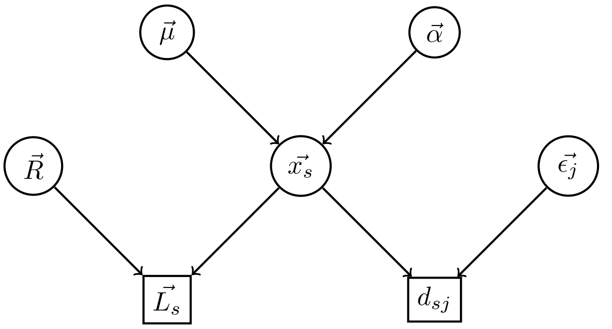

Figure 2: Potential DAG of the model.

With $P$ the phenotype between two species, $E$ an environment factor between two species and $M$ the TODO

# 3 Formulation

We seek to infer the presence or absence of $M$ metabolites in tissue $T$ in $S$ species. We denote by $x_{smt}$ whether metabolite $m = 1, \dots, M$ is present ($x_{smt} = 1$) or absent ($x_{smt} = 0$) in tissue $t = 1, \dots T$ in species $s = 1, \dots, S$. We denote $\vec{x_{st}} = (x_{st1}, \dots, x_{stM})$ the vector of molecules present in a tissue $T$ of a specific species $S$.

Let us further denote $\vec{x_s} = (\vec{x_{s1}}, \dots, \vec{x_{sT}}) = (x_{s11}, \dots, x_{s1M}, x_{s21}, \dots, x_{sTM})$ the vector of presence/absence of all molecules across all tissues for species $s$.

We have two origins of data, mass spectrometry data and the Lotus database. 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 1. $\vec{\mu}$ being the vector of the average presence/absence of each molecule across all species and $\vec{\alpha}$ represents the vector of all known measures of covariates between species and molecules.

Since the tissue specific origin of a molecule is not known in the LOTUS database [1], we denote

$$P(\vec{L_S}|\vec{x_s}, R) = P(\vec{L_s}|\vec{\xi_s}, \vec{R}),$$

with $\vec{\xi_s} = (\xi_{s1}, \dots, \xi_{sM})$ the vector of presence/absence of all molecules $M$ in species $s$. Furthermore, $\xi_{sm} = min(1, \sum_t^T x_{smt})$ the minimum between 1 and the sum of presence or absence of a molecule across all tissues.

We can also denote $\vec{R} = (R_{11}, \dots, R_{1M}, R_{21}, \dots, R_{SM})$ the vector of *research effort* of all molecules across all species.

The probability of having a molecule present in the LOTUS database not only depends on the presence/absence of that molecule in a species but also on the research effort done for a specific molecule or species. We thus have $R_{sm} = f(n_s, n_m)$ with $R_{sm} \in [0, 1]$ and where $n_s$ and $n_m$ are the number of scientific papers that relate respectively the species or the molecules of interest. We thus have the following matrix :

$$\begin{array}{c|cc} & L_{sm} = NA & L_{sm} = 1 \\ \hline x_{sm} = 0 & 1 & 0 \\ x_{sm} = 1 & 1 - R_{sm} & R_{sm} \end{array}$$

Tissue of origin is usually known in mass spectrometry analysis. From Figure 1, we have $P(d_{sj}|\vec{x_s}, \vec{\epsilon_j}) = P(d_{sj}|x_{st(\vec{d_{sj}})}, \vec{\epsilon_j})$ where $t(d_{sj})$ reflects the tissue from which mass spectrometry run $j$ in species $s$ was sampled.

We thus have:

$$P(\boldsymbol{d}, \boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\alpha}) = P(\boldsymbol{\mu})P(\boldsymbol{\alpha}) \prod_{s=1}^{S} P(\vec{x_s}|\boldsymbol{\mu}, \boldsymbol{\alpha}) \prod_{j=1}^{J} P(d_{sj}|x_{st(\vec{d_{sj}})}, \vec{\epsilon_j}) \tag{4}$$

Similarly, for LOTUS database we have :

$$P(\text{Ł}, \boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\alpha}) = P(\boldsymbol{\mu})P(\boldsymbol{\alpha})P(\mathbf{R}) \prod_{s=1}^{S} P(\vec{x_s}|\boldsymbol{\mu}, \boldsymbol{\alpha})P(L_s|\vec{x_s}, \mathbf{R}) \tag{5}$$

### 3.1 Decreasing complexity

If one is interested in the presence/absence of a molecule not in the species but on the Genus/Order level, one can easily remodel the previous model such as $\vec{x_g} = (\vec{x_{g1}}, \ldots, \vec{x_{gM}})$ and where $\vec{x_{gm}} = min(1, \sum_{s \in g} \sum_{t}^{T} x_{smt})$.

## 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. `doi:10.7554/eLife.70780`.