Inferring the presence of metabolites

March 21, 2023

1 Model

1.1 Definition

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 $\vec{w} = \vec{v} \setminus \{\vec{s}, \vec{m}\}$ of length K with $K = I - 2$. Similarly with $\vec{v}$ we have $\vec{w} = \{\vec{w_1}, \ldots, \vec{w_K}\}$ with $\vec{w}_k = \{w_1, \ldots, w_{L_k}\}$.

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

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

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

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

with I the length of vector $\vec{v}$ (or $\vec{c}$), 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.

1.1.1 Updating parameters

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 :

$$P(\vec{\alpha}|\vec{x_s}, \vec{\mu}) \propto P(\vec{x_s}|\vec{\mu}, \vec{\alpha})P(\vec{\alpha}) \tag{3}$$

and,

$$P(\vec{\mu}|\vec{x_s}, \vec{\alpha}) \propto P(\vec{x_s}|\vec{\mu}, \vec{\alpha})P(\vec{\mu}) \tag{4}$$

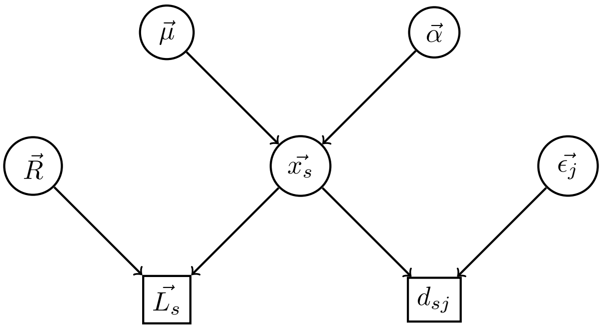

Figure 1: Potential DAG of the model.

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

$$q_\alpha = min\left[1, \frac{P(\vec{x_s}|\vec{\mu}, \vec{\alpha'})}{P(\vec{x_s}|\vec{\mu}, \vec{\alpha})} \cdot \frac{P(\vec{\alpha'})}{P(\vec{\alpha})}\right] \quad \text{and,} \quad q_\mu = min\left[1, \frac{P(\vec{x_s}|\vec{\mu'}, \vec{\alpha})}{P(\vec{x_s}|\vec{\mu}, \vec{\alpha})} \cdot \frac{P(\vec{\mu'})}{P(\vec{\mu})}\right] \tag{5}$$

Furthermore, 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 1.

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 $\vec{\xi_s} = (\xi_{s1}, \ldots, \xi_{sM})$ 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.

Based on the defined vectors in Section 1.1, we can assume a more general case with :

$$\xi_{sm} = min\left[1, \sum_{k=1}^{K}\sum_{l=1}^{L_k} x_{smw_{k_l}}\right] \quad , \tag{6}$$

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 :

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

1.3 MS data

Let $\vec{d_{sj}} = (d_{sj1}, \ldots, d_{sjM})$ be the presence-absence vector of each metabolite m obtained with mass-spectrometry run $j = 1, \ldots, J_s$ performed on species s. Assuming a false-positive and false-negative error rates ϵ_{01} and ϵ_{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 v2

We seek to infer the presence or absence of metabolites in group of samples compartmentalized by an arbitrary number of discrete axis such as e.g. species, tissue or environmental conditions. Let $\boldsymbol{\tau} = \{\tau_1, \ldots, \tau_T\}$ denote the set of T axis of compartmentalization and let, without loss of generality, $\tau_1 = \mathcal{M}$ be the axis of metabolites. For any compartment c, let $\tau_t(c) = 1, \ldots, n_t$ indicate the compartment index along axis τ_t with $\tau_\mathcal{M}(c) = \tau_1(c)$ indicating the metabolite of that compartment.

Let x_c denote the presence ($x_c = 1$) or absence ($x_c = 0$) of a metabolite $\tau_\mathcal{M}(c)$ in compartment c and let $\boldsymbol{x} = (x_1, \ldots, x_C)$ be the full vector x_c across all compartments $c = 1, \ldots, C$.

We will assume that similarities across any of the axis of compartmentalization is reflected in the patterns of presences and absences in $\boldsymbol{x}$. 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 assume that the probability $\mathbb{P}(x_c = 1|\boldsymbol{\mu}_c, \epsilon_c)$ with which metabolite $\tau_\mathcal{M}(c) = m$ is present in compartment c is given by

$$\operatorname{logit} \mathbb{P}(x_c = 1|\boldsymbol{\mu}_c, \epsilon_c) = \sum_{t=1}^{T} \mu_t + \epsilon_c \tag{7}$$

where $\boldsymbol{\mu}_c = (\mu_1, \ldots, \mu_T)$ is a vector of axis specific intercepts and ϵ_c is normally distributed with mean 0 and co-variance

$$\operatorname{cov}(\epsilon_c, \epsilon_{c'}) = \sum_t \beta^{(\tau_t)}_{\tau_t(c)} + \sum_t \beta^{(\tau_t)}_{\tau_t(c')} + \sum_f \alpha_f \sigma_f \Big(\tau_{\tau(f)}(c), \tau_{\tau(f)}(c') \Big). \tag{8}$$

Here, the $\beta^{\tau_t}_{\tau_t(c)}$ are intercepts specific for the compartment index $\tau_t(c)$ along axis τ_t, the $\sigma_f, f = 1, \ldots, F$, are known covariances between entries along axis $\tau(f)$ and the α_f are positive scalars.

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`.