Phylogenic Trees

This version 2017-01-18

Definitions

A set of genes $N \equiv \{n \in \{1, \dots, G\}\}$ is a partial order according to its relations as parent and offspring such that for any $n, m \in N$ for which a directed path exists, $n < m$ iff n is a parent node and m an offspring.

Leaf Probabilities

Given $1, \dots, P$ functions, each node has 2^P different possible states. For any given node n, its state s_n is a vector of length P in $\{0,1\}^P$, *e.g.* $s_n = \{0, 0, 0\}$.

In the experimental data, for each leaf l, we have the observed state defined by the vector $s_l \equiv \{s_{lp}\}_{p=1}^P$ with

$$z_{lp} = \begin{cases} 1 & \text{if the function } p \text{ is active} \\ 0 & \text{if the function } p \text{ is not active} \\ 9 & \text{if we don't have information} \end{cases}$$

This way, given that we observe $s' \equiv \{s'_p\}_{p=1}^P$, the probability that the true state is s is:

$$\Psi_{ss'} = \begin{cases} \prod_p \left\{ \left[\psi_0^{s'_{lp}} (1-\psi_0)^{\neg s'_{lp}}\right]^{\mathbf{1}\{\neg s_{lp}\}} \left[\psi_1^{\neg s'_{lp}} (1-\psi_1)^{s'_{lp}}\right]^{\mathbf{1}\{s_{lp}\}} \right\} & \text{if } s'_{lp} \neq 9 \\ & \\ 1 & \text{otherwise} \end{cases}$$

Where $\psi \equiv \{\psi_0, \psi_1\}$ is a vector of missclassification probabilities for un-activated and activated functions.

Internal Nodes Probabilities

For any internal node n, the likelihood is defined in terms of gain and loss functions (also stored as a $N \times 2^P$ array). Furthermore, it is conditional on n's offspring $o_n \subset \{m \in N : m < n\}$, which has cardinality $|o_n| = O_n$, and the true state s. Then, the probability that the internal node n has state s is

$$P_{n,s} = \prod_{o_n} \sum_{s^*} P_{o_n,s^*} \prod_p \left(\left[\underbrace{\mu_0^{\mathbf{1}\{s_p^*\}}}_{\text{Gain}} \underbrace{(1-\mu_0)^{\mathbf{1}\{\neg s_p^*\}}}_{\text{No gain}} \right]^{\mathbf{1}\{\neg s_p\}} \times \left[\underbrace{\mu_1^{\mathbf{1}\{\neg s_p^*\}}}_{\text{Loss}} \underbrace{(1-\mu_1)^{\mathbf{1}\{s_p^*\}}}_{\text{No Loss}} \right]^{\mathbf{1}\{s_p\}} \right)$$

Where $P_{o_n,s^*} = \Psi_{s^*s'}$ if the offspring is a leaf. Computationally, observe that the larger parenthesis can be computed only once and then retrieved depending on the values of $\{s_p^*, s_p\}$. Let $M \equiv \{m_{s_p^*,s_p}\}$ to be an array of size 2×2 holding the Gain/Loss probabilities, then, the previous equation reduces to:

$$P_{n,s} = \prod_{o_n} \sum_{s^*} P_{o_n,s^*} \prod_p m_{s_p^*,s_p}$$

Finally, let $\pi \equiv \{\pi_s\}_{s=1}^{2^P}$ to be the root node state probabilities, then, the likelihood for $n = 0$ can be computed as

$$L_0(\pi, \mu, \psi) = \sum_s \pi_s P_{0,s}$$

Data

The model uses two different datasets: (1) experimental data, which holds functions indicators, and (2) phylogenetic tree data, which contains the parent/offspring relations.

Given how the algorithm for computing the likelihood has been programmed, it is necessary that the experimental dataset must have as many rows as nodes (parents and offspring) there are. To fulfill such requirement, the package `phylogenetic`, throught the function `prepare_data`, "completes" the experimental dataset as follows:

1. List all nodes in the tree that are not in the experimental dataset.
2. Once identified, if any, add those nodes to the experimental dataset. The function indicator columns will have value '9' (unknown). Furthermore, the added nodes are tagged so that the user can identify them later.
3. The new experimental dataset is sorted increasingly according to the node id number. The idea is that the root, node 0, should appear in the first row.

Once the experimental data has been processed,

4. We identify