1 Detailed Methods

Arboretum is a model-based clustering approach that uses a probabilistic generative model to cluster multiple expression datasets, one for each species. Each dataset resides at a leaf node of a species tree describing the relationships between the species. Arboretum works with a fixed, sufficiently large number of clusters, and assumes that if there are different numbers of clusters in different species, those clusters will be empty.

The generative model has to generate values for two types of random variables: (a) hidden variables representing the cluster assignments in both ancestral and extant species, and (b) observed variables encoding expression for each gene in a species. The cluster membership is modeled by conditional distributions for every branch of the species tree, describing the probability of a gene belonging to a cluster in a species given the cluster membership in its immediate ancestor. The expression data at each leaf node is modeled by a Gaussian mixture model. An integral part of Arboretum is that it naturally handles one-to-many mappings of genes over any number of species. This is done by incorporating the gene tree directly inside Arboretum's cluster inference. In the following sections we describe the different parts of the model in detail, inference of cluster assignments and parameter estimation.

1.1 Modeling cluster assignments and their evolution

Let K denote the maximum number of clusters that can exist in a species. We assume that every gene in all the extant species evolve their cluster assignment from a single ancestral version which is present at the most ancient ancestor (root of the species tree). The most ancient ancestor (last common ancestor) has a prior probability distribution, a multinomial, $p(k)$, $1 \leq k \leq K$, which specifies an initial assignment to a cluster for a gene. Every other species t has a cluster transition matrix, P_t which specifies a conditional probability distribution of cluster assignments of genes in t, from cluster assignments of genes in t's immediate ancestor. In particular, every element in the transition matrix $P_t(i, j)$ is the conditional probability of a gene to be in cluster i given that its ancestral version was in cluster j.

We refer to a set of orthologous genes as an orthogroup. Every orthogroup has an associated gene tree, which species the phylogenetic relationships between the genes in the orthogroup. An orthogroup which has one-to-one mapping across species is called a *uniform* orthogroup. Because every species has one and only representation of a gene, the corresponding gene tree for these orthogroup is the same as the species tree. The cluster evolution process generates the cluster assignment of all genes in an orthogroup, at a time, by sampling a cluster assignment from a prior multinomial distribution at the root, propagating the assignment down the tree via the transition matrices along the branches of the species tree. At the leaf nodes we generate expression of a gene from the Gaussian indicated by the propagated cluster assignment. Thus the evolution process takes into account the phylogenetic relationships across the species. The cluster assignments of all the species are related to each other using the species tree. We use this tree structure to devise a tractable cluster inference procedure (Section 1.3.1).

1.1.1 Modeling cluster assignments in non-uniform orthogroups

A non-uniform orthogroup is one where genes may have undergone a duplication event (Fig 1) in a subset of the species, because of which there is no longer a one-to-one mapping across species. For these orthogroups, the gene tree also describes duplication and loss events and thus the structure of the gene tree is different from the species tree. To handle duplicates, we proceed down the tree as in the usual uniform orthogroup case until we reach a species where a duplication happened. At the duplication node, we draw two samples from the cluster transition probability matrix, each of which can then be evolved down the rest of the tree independently following the same procedure as before. This also captures the fact the paralogous genes often evolve along different trajectories allowing one gene copy to possibly acquire new function.

1.2 Modeling observed expression data

Let $\mathbf{X}_t$ denote the set of random variables encoding the expression values in species t. $\mathbf{x}_i$ denotes the expression profiles of the genes associated with the extant species of the i^{th} orthogroup. The expression data at species t is

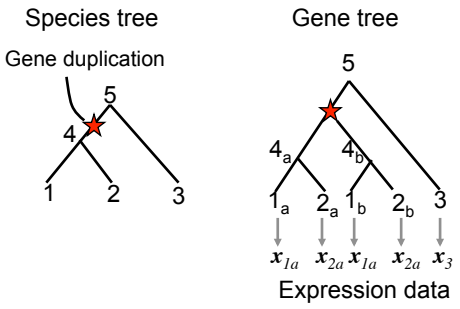

Figure 1: Species and gene tree with a single duplication event

modeled by a mixture of K Gaussians, where the K is the total number of clusters:

$$\mathbf{x}_{ti} \sim \sum_{k=1}^{K} \alpha_k \mathcal{N}(\mu_{tk}, \Sigma_{tk})$$

where the k^{th} mixture component models the expression profile of the k^{th} module. μ_{tk} is a D_t-dimensional mean vector where D_t is the number of measurements available for each gene in species t. Thus, Arboretum can be easily applied to datasets with different values of D_t, which is the common case for comparative functional datasets as not all species may be measured as frequently as other species.

1.3 Expectation Maximization framework for learning

We use the expectation maximization (EM) framework to infer the posterior probability distribution of these cluster assignments. There are two steps in this framework: expectation, during which hidden variables are inferred from the current model parameters, and maximization, during which parameters are estimated from the expected values of the hidden variables.

1.3.1 Expectation: Inference of cluster assignments

Let $\mathbf{Z}$ denote the set of hidden variables specifying the cluster assignments. Let θ denote the set of parameters associated with the model. Let $\mathbf{z}_i$ is a set of hidden variables for all genes associated with the i^{th} orthogroup. Let $2N-1$ denote the total number of species, and thus N is the number of extant species. We refer to the extant species using the index t, $1 \leq t \leq N$, and the ancestral species using $N+1 \leq t \leq 2N-1$. The root of the tree is referred by the index $r = 2N-1$.

Let τ_i denote the tree associated with the i^{th} orthogroup.The hidden variables in each orthogroup are related to each other via τ_i's structure. At the root of the tree, i.e. at the LCA, z_{ri}, denotes the cluster assignment of the most ancestral version of a gene in this orthogroup. z_{ri} is a multinomial random variable. For all other nodes, $1 \leq t < 2N$, $z_{ti}^{k|k'}$ denotes the conditional cluster membership of t's gene in cluster k, given that its ancestral version is in k'. Our inference problem is to infer the posterior probability of these hidden variables given the data, $P(\mathbf{z}_i|\mathbf{x}_i)$, where $\mathbf{x}_i$ denotes the expression profiles of the genes associated with the extant species.

Let $\mathbf{x}_{ti}$ be the subset of the expression data from the extant species that are under the t^{th} node. Let $\gamma_{ti}^{k|k'}$ be the posterior probability for t's gene to be in cluster k, given that the immediate ancestral version of this gene is in cluster k', i.e. $P(z_{ti}^{k|k'}|\mathbf{x}_i)$. To infer this posterior probablity, we make a crucial independence assumption that allows us to perform tractable inference: we assume that the cluster assignment of a gene is dependent only upon the subset of the expression data that comes from the subtree below it. This assumption is the similar to the one made in Felsenstein's Pruning algorithm used to infer the likelihood of a given DNA sequence. Thus $P(z_{ti}^{k|k'}|\mathbf{x}_i) = P(z_{ti}^{k|k'}|\mathbf{x}_{ti})$. This allows us to compute the posterior probability at each internal node using computation from it's child nodes as we show below.

Demonstrative example for inference of cluster assignments We consider an illustrative example of 3 extant species indexed by $\{1, 2, 3\}$ and 2 ancestral species, $\{4, 5\}$ where 4 is common ancestor of 1 and 2 and 5 is the common ancestor of 4 and 3. Also, for simplifying notation, we focus on a particular orthogroup and therefore drop i. Thus $\mathbf{z}_i = \mathbf{z}$, $\mathbf{x}_i = \mathbf{x}$, $\mathbf{x}_{ti} = \mathbf{x}_t$ etc.

At the root, $t = 5$, we have, $\gamma_5^k = P(z_5^k|\mathbf{x}_5) = \frac{P(z_5^k)P(\mathbf{x}_5|z_5^k)}{P(x_5)}$. Note, $\mathbf{x}_5 = \mathbf{x}$, that is, all the expression data associated with the orthogroup in question. $P(\mathbf{x}_5|z_k^5)$ is the likelihood of $\mathbf{x}_5$ given the model, and can be obtained as: $P(\mathbf{x}_5|z_5^k) = \sum_{z_4^{k|k'}} \sum_{z_3^{k|k'}} P(\mathbf{x}_3, \mathbf{x}_4, z_4^{k|k'} z_3^{k|k'}|z_5^{k|k'})$. Assuming our tree-based independence, this can re-written as $\sum_{z_4^{k|k'}} \sum_{z_3^{k|k'}} P(\mathbf{x}_3|z_3^{k|k'})P(z_3^{k|k'}|z_5^k)P(\mathbf{x}_4|z_4^{k|k'})P(z_3^{k|k'}|z_5^k)$, which can be re-arranged into a product

$$\left(\sum_{z_4^{k|k'}} P(\mathbf{x}_4|z_4^{k|k'})P(z_4^{k|k'}|z_5^k)\right)\left(\sum_{z_3^{k|k'}} P(\mathbf{x}_3|z_3^{k|k'})P(z_3^{k|k'}|z_5^k)\right).$$

The conditionals $P(z_3^{k|k'}|z_5^k)$ are nothing but the prior probability of a gene being in a cluster k in species $t = 3$ given that in ancestor 5, the gene is in cluster k'.

Similarly, at $t = 4$, $\gamma_4^{k|k'} = P(z_4^{k|k'}|\mathbf{x}_4) = \frac{P(\mathbf{x}_4|z_4^{k|k'})P(z_4^{k|k'}|z_5^k)}{\sum_{z_4^{k|k'}} P(\mathbf{x}_4|z_4^{k|k'})P(z_4^{k|k'}|z_5^k)}$. Note, the normalization term is the same as the first term of the numerator from γ_5^k. Similar if we expanded $\gamma_3^{k|k'}$, the denominator is the same as the second term of the numerator. As in the case of γ_5^k, $P(\mathbf{x}_4|z_4^{k|k'}) = \sum_{z_1^{k|k'}} \sum_{z_2^{k|k'}} P(\mathbf{x}_1, \mathbf{x}_2, z_1^{k|k'}, z_2^{k|k'}|z_4^{k|k'})$. Finally, the posterior probability of the leaf nodes $\gamma_3^{k|k'}, \gamma_2^{k|k'}, \gamma_1^{k|k'}$ can be easily computed using the data at that node, e.g.

$$\gamma_3^{k|k'} = \frac{P(\mathbf{x}_3|z_3^{k|k'})P(z_3^{k|k'}|z_5^{k|k'})}{\sum_{z_3^{k|k'}} P(x_3|z_3^{k|k'})P(z_3^{k|k'})}$$

The normalization constants estimated at each child node is re-used in the estimation of the posterior probability distribution at the parent parent node.

Inference of cluster assignments of orthogroups with duplicates Let us now consider an example of an orthogroup with a duplication. We assume that a single duplication occurred at ancestor $t = 4$, which is the parent of 1 and 2. We denote this duplication as a branching event on the tree (Fig 1), after which we will have two copies in the node $t = 4$, and also in species 1 and 2. Let the duplicate expression for these species be denoted as $\mathbf{x}_4 = \{\mathbf{x}_{4a}, \mathbf{x}_{4b}\}$, $\mathbf{x}_{4a} = \{\mathbf{x}_{1a}, \mathbf{x}_{2a}\}$ and $\mathbf{x}_{4b} = \{\mathbf{x}_{1b}, \mathbf{x}_{2b}\}$. Because we assume that after duplication the two duplicates evolve independently, the contribution from the sub-tree below the duplication is simply a product,

$$P(x_4|z_k^5) = \left(\sum_{z_{4a}^{k|k'}} P(\mathbf{x}_{4a}|z_{4a}^{k|k'})P(z_{4a}^{k|k'}|z_k^5)\right)\left(\sum_{z_{4b}^{k|k'}} P(\mathbf{x}_{4b}|z_{4b}^{k|k'})P(z_{4b}^{k|k'}|z_5^k)\right).$$

In general, our recursive inference procedure relies on the observation that the computation we perform at a non-root node, t to estimate the posterior probability at that node is used for estimating the normalization constant t's parent. We begin at the leaf nodes to estimate the γ's. The product of γ's at two sibling leaf nodes would then give the normalization constant for the intermediate node. Subsequently we would obtain the normalization constants of an intermediate node by taking the product of its subtrees. When we reach the root node, the product of the subtrees give the full posterior distribution of the joint of cluster assignments given the expression data pertinent to the orthogroup.

1.4 Maximization: estimation of parameters

There are two sets of parameters in this model: (a) cluster transition probabilities, (b) Gaussian mixture model parameters. We assume that the co-variance matrix is a diagonal matrix. Each of these parameters can be estimated in closed

form by deriving the expected likelihood with respect to the parameters. The expected likelihood in turn is taken as a sum over all orthogroups, $\mathcal{G}$. Thus our likelihood is written as:

$$Q(\theta, \theta') = \sum_{i}^{|\mathcal{G}|} \sum_{\mathbf{z}_i} \Gamma_i \log P(\mathbf{x}_i|\mathbf{z}_i) P(\mathbf{z}_i)$$

Γ_i in turn is written as a product: $\gamma_{ri}^{k} \prod_{t=1}^{r-1} \gamma_{ti}^{k|k'}$, where t is the species index and r is the total number of species. This sum over products can be re-arranged to yield a sum of two groups of terms: those that involve the expression data $\mathbf{x}_i$, and those that do not:

$$\begin{aligned} Q(\theta, \theta') = & \sum_{i}^{|\mathcal{G}|} \sum_{z_{ri}^{k}} \gamma_{k}^{ri} \log P(z_{ri}^{k}) \\ & + \sum_{z_{si}^{k|k'}} \gamma_{si}^{k|k'} \log P(z_{si}^{k|k'}) + \cdots + \sum_{z_{ti}^{k|k'}} \gamma_{ti}^{k|k'} \log P(z_{ti}^{k|k'}) \\ & + \sum_{z_{qi}^{k|k'}} \gamma_{qi}^{k|k'} \log P(z_{qi}^{k|k'}) P(\mathbf{x}_{qi}|z_{qi}^{k|k'}) + \cdots + \sum_{z_{pi}^{k|k'}} \gamma_{pi}^{k|k'} \log P(z_{pi}^{k|k'}) P(\mathbf{x}_{pi}|z_{pi}^{k|k'}) \end{aligned} \tag{1}$$

$$\tag{2}$$

Here r represents the root species, s to t are all the non-leaf nodes other than the root, and q to p are all the leaf nodes with expression data avaliable. The maximum likelihood mean estimate for j^{th} cluster for the t^{th} species, μ_{jt} is very similar to the standard Gaussian mixture model case, except the hidden variables that come into play $\gamma_{k}^{t|t'}$, takes k^2 values rather than k values:

$$\mu_{jt} = \frac{\sum_i \sum_{l=1}^{k} \gamma_t^{j|l} \mathbf{x}_{ji}}{\sum_i \sum_{l=1}^{k} \gamma_t^{j|l}}$$

Similarly for the variance estimate, we need an additional sum to account for the fact that the cluster assignment in an extant species is dependent upon its parents.

The transition probabilities for each species is estimated from the expected value of the joint assignment of a child and parent cluster assignment pair, $P(z_{ti} = k, z_{ui} = k'|\mathbf{x}_i)$, which is $P(z_{ti} = k|z_{ui} = k', \mathbf{x}_i) P(z_{ui} = k'|\mathbf{x}_i)$. Note $P(z_{ti} = k|z_{ui} = k', \mathbf{x}_i) = \gamma_{ti}^{k|k'}$. To obtain the marginal $P(z_{ti} = k|\mathbf{x}_i)$, we begin with the root r for which we already have the estimated probability values. Then we simply descend from the root down one level where we estimate $P(z_{ti} = k, z_{ri} = k'|\mathbf{x}_i)$ by multiplying $P(z_{ti} = k|z_{ri} = k', \mathbf{x}_i)$ with the marginal of its parent r. From this joint we can get the marginal $P(z_{ti} = k|\mathbf{x}_i)$, which can be used for the children of t. In this way we recursively estimate the joint and marginals at each node. Then the $P_t(k, k')^{th}$ entry is given by $\frac{\sum_i P(z_{ti}=k, z_{ui}=k'|\mathbf{x}_i)}{\sum_i P(z_{ui}=k'|\mathbf{x}_i)}$. If an orthogroup has two genes for a species, both of them contribute to $P_t(k, k')$.

1.4.1 Learning algorithm

Now that we have all the parts of the model, we can describe our learning algorithm. We begin with an initial clustering assignment obtained from partitioning all orthogroups into K partitions, where K specifies the number of clusters per species. This partitioning can be obtained by randomly splitting the data, or by a clustering algorithm that merges all the species data together into a single vector and clusters these concatenated data. The clustering is not expected to be good because orthologous genes may not cluster together. Further, the genes with many to one mappings would need their data to be duplicated for the columns corresponding to the species with one copy of a gene.

Because the EM algorithm may permute the cluster indices during learning, we take two measures. The initializations of the transition matrices have heavy diagonals, i.e., the probability of a species to conserve a gene's cluster assignment from its immediate ancestor is high. We also have two rounds of EM. After the first EM training, we check for cases where the cluster assignments is conserved in all intermediate nodes from a leaf to the root, except at the

leaf node. If such a case arises, we swap the probabilities of a gene belonging to the cluster at the leaf and the rest of the path to the root, and perform another round of EM. The algorithm uses these initial cluster partitions to seed the parameters values for the Gaussian mixtures.

Algorithm 1 Learning in Arboretum

Input:
Number of clusters k
Datasets for $s \in \mathcal{S}$, $\{\mathcal{D}_1, \cdots, \mathcal{D}_{|\mathcal{S}|}\}$
A partition of datasets into k clusters
Initial assignment to transition matrix
Species tree τ_s
Gene trees τ_i for each orthogroup O_i
Output:
Inferred cluster assignments for both extant and ancestral species
Clustered expression data
A set of transition probabilities of cluster assignments
Initialize parameters for K Gaussian distributions for each species using the input partition.
while Likelihood does not stabilize **do**
/*Expectation Step*/
Infer the expected values of cluster assignments, $\gamma^t_{k|k'}$ of all genes at the leaf nodes.
Recursively infer the expected values of cluster assignments of an intermediate node using the data nodes in the subtree below it.
/* Maximization Step*/
Estimate $\mu_{j,s}$ and $\Sigma_{j,s}$ using $\gamma^t_{k|k'}$ for all species t at the leaf nodes.
Estimate T_s for all species $s \in \mathcal{S}$.
end while