© 2016 IJSRSET | Volume 2 | Issue 5 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099
Themed Section: Engineering and Technology

Constructing a Fully Ranked phylogenetic Constraint Tree using Monophyletic Group, Crown Group and Relative Age Constraints

Aiasha Siddika[1], Md. Towhidul Islam Robin[2], Umme Rumman Usha[2]
[1,2]Stamford University Bangladesh, Dhaka, Bangladesh
[3]Bangladesh University of Engineering and Technology, Dhaka, Bangladesh

ABSTRACT

Phylogenetic research field concerned with the reconstructing of the phylogenetic tree using some information that has evolved from the root to all the descendants. There are many ways to try and find the best tree.If one knows beforehand that a subset of species form a subgroup that are closely related then one can constrain the search algorithm to only look for those trees where this subset of species forms a subtree. Placing such constraints reduces the number of trees that the search algorithm must consider and thus reduces the time spent searching for the best tree. In Bayesian phylogenetic inference we are interested in distributions over a space of trees. When fossil evidence is used in the inference to constrain the tree, new tree spaces arise and counting the number of trees is more difficult. We have constructed a tree using algorithm that is polynomial in the number of sampled individuals for counting of resolutions of a constraint tree assuming that the number of constraints is three.

Keywords: Phylogenetics, Ranked Tree, Tree Counting, Dynamic Algorithms, Bayesian Inference, Full Constraint Tree.

I. INTRODUCTION

Phylogenetics is the study of evolutionary relationships among biological entities (often species, individuals or genes which may be referred to as taxa). The tree that represents such relationships among species is called phylogenetic tree. The leaves of the tree represent existing species, the internal vertices represent ancestors, the edges represent evolutionary steps, the root represents the oldest evolutionary ancestor of the existing species represented in the tree. Point to be noted that phylogenetic trees can be either rooted (if the common ancestor is known) or unrooted (if the common ancestor is unknown) [9].

Now-a-days a phylogenetic tree is the common object of interest in many areas of biological as well as computational science. Given molecular sequence data sampled from a group of organisms it is possible to infer the historical relationships between these organisms using a statistical model of molecular evolution. At present, Bayesian Markov chain Monte Carlo (MCMC) methods are the dominant inferential tool for inferring molecular phylogenies [10].

When fossils are used to restrict the age of internal nodes, the tree prior should accurately account for this fact. Heled and Drummond [6] introduced a natural approach for tree prior specification when fossil evidence is employed in the inference. Their method requires counting of ranked phylogenetic trees that obey a number of constraints that arise from including the fossil evidence. The construction requires calculation of the marginal density for the time of the calibration node, the node representing the most recent common ancestor of a clade which may or may not be monophyletic. For a particular location of the calibration node, or particular constraints on the tree topology, the marginal density function is the marginal density function for the divergence times weighted by the number of trees satisfying the constraints. In this case, the weight constants do not cancel in the MCMC scheme and therefore have to be calculated.

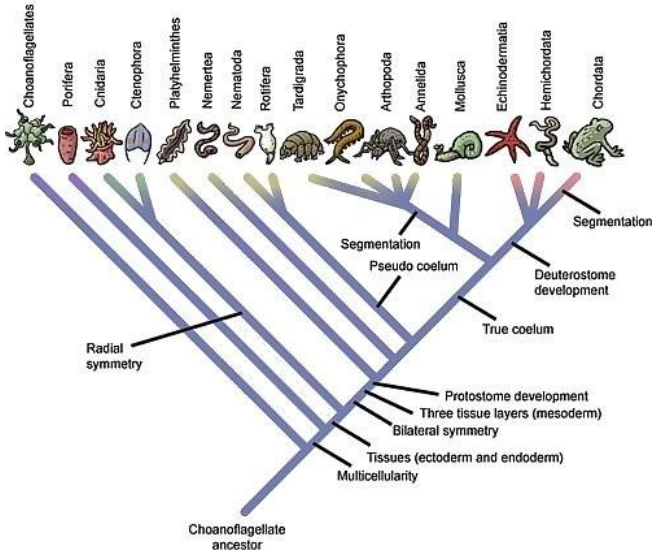

Figure 1: Phylogenetic tree of the Animal kingdom [2].

For phylogenetic trees, the tree counting problem is to find the number of all possible trees on n leaves. For some types of phylogenetic trees, there are known closed form solutions to this problem. For other types, only recursive equations have been derived. Using such recursive algorithms, we can generate several types of phylogenetic trees, such that: Ranked X-trees, Fully Ranked X-trees, Fully Ranked X-trees with sampled ancestors, Constraint X-tree, FRS Constraint X-tree etc [1].

Tree counting has a long history. For phylogenetic trees, the counting problem is to find the number of all possible trees on n leaves. For some types of phylogenetic tree, there are known closed form solutions to this problem. For other types, only recursive equations have been derived. The number of trees in a tree space is an important characteristic of the space and is useful for specifying prior distributions. When all samples come from the same time point and no prior information available on divergence times, the tree counting problem is easy. However, when fossil evidence is used in the inference to constrain the tree or data are sampled serially, new tree spaces arise and counting the number of trees is more difficult. A survey of results on counting different types of rooted trees is presented in [3] where trees with different combinations of the following properties are considered: trees are either labeled (only leaves are labeled) or unlabeled, ranked or non-ranked, and bifurcating or multifurcating. The results presented in the survey can also be found in [8].

In [5], Felsenstein considered partially labeled trees, i.e., a tree in which all the leaves are labeled and some interior nodes also may be labeled. He derived the recursive equations for counting the number of rooted, non-ranked, partially labeled trees with n labeled nodes. Complexity of tree counting of different variations of the extended problem has been analyzed in these recent years. Given an edge-weighted tree TT with leaf set X, define the weight of a subset S of X as the sum of the edge-weights of the minimal subtree of TT connecting the elements in S. It is known that the problem of selecting subsets of X of a given size to maximize this weight can be solved using a greedy algorithm. This optimization problem arises in conservation biology where the weight is referred to as the phylogenetic diversity of a taxa set S. Here, we consider the extension of this problem whereby we are only interested in selecting subsets of the taxa set that are ecologically "viable". Such subsets are specified by an acyclic digraph which represents, for example, a food web. This additional constraint makes the problem computationally hard.[4]

In our report, we will focus on constraint trees only. In phylogenetic analysis, it is common to have some limited information about the ancestors of sampled individuals, such information acts as constraints for constructing phylogenetic trees. Here, we consider three such constrains: monophyletic, crown group [7] and relative age.

A subgroup of some sampled individuals is called monophyletic if the most recent common ancestor of that subgroup is not an ancestor of any other individual that does not belong to the subgroup. Crown group is basically a clade or group consisting of a species and all its descendants. It is quite possible to know the relative ages of the most recent common ancestors of monophyletic subgroups [9] or crown groups which are usually presented as the rankings of the constraint X-tree.

Counting of resolutions of a constraint tree is an expensive procedure. But if the number of constraints, k is small, the tree construction is quite feasible. In practice, k is typically small, and in our case k=3, so this algorithm will be of practical use.

II. METHODS AND MATERIAL

Let us consider a phylogenetic tree T. Since the relative ages of the species of T are known, we can say that T is a ranked tree. These ranks will be determined using a ranking function, r. So, if there are two vertices (which can be either interior or leaf) v1 and v2; then v1 <= T v2 implies that r (v_1) <= r (v_2). A node in a rooted tree is called interior if it has descendants and a leaf if it has no descendants. The root is considered interior. Rooted tree is called binary if every interior node has exactly two children. It is called weakly binary if every interior node has at most two children.

We consider T to be a rooted phylogenetic tree. This T will represent the relationships among a set of species, let the set be X = X_1, X_2, X_3, . . . , X_n. So, |X| = n. Also we will count on some constrains like monophyletic or crown group, for which the prior information of the common ancestors should be known.

A Ranked X-tree is the simplest version of such phylogenetic tree. When all individuals are sampled at the same time, a simple Ranked X-tree is generated while solving the counting tree problem. A figure of such a Ranked X-tree is given in figure 1.

In biology, a phylogenetic tree represents the evolutionary history of a collection of sampled individuals. The collection of individuals is represented by the set X. The root of the tree is the most recent common ancestor of X and interior nodes are bifurcation events. The ranking function represents the time order of the bifurcation events. A general problem in evolutionary biology is how to reconstruct the phylogenetic tree from sequence data obtained from sampled individuals. Tackling this problem in a Bayesian framework may require counting the number of all possible histories on a sample of individuals. When all individuals are sampled at the same time (as in Figure 1) counting tree problem has a simple solution.

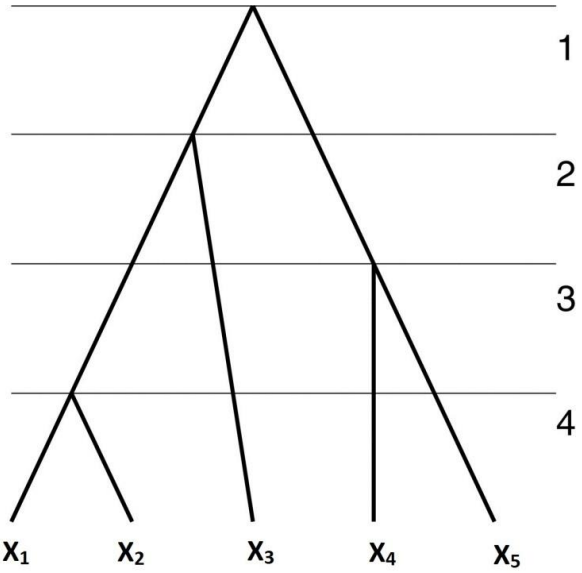

Figure 2: Ranked X-tree, $X = X_1, X_2, X_3, X_4, X_5$. The numbers on the right are values of the ranking function. [1]

The number of all ranked X-trees up to isomorphism is

$$R(n) = \frac{n!\,(n-1)!}{2^{(n-1)}}$$

Here, the letter R in the equation comes from the word "Ranked". This formula has been derived by many authors. Proofs can be found in [3, 8 and 4].

This scenario can be different if the individuals are serially sampled at different times. To represent such relationship, we can construct a fully ranked (FR) X-tree. Before the tree is reconstructed we observe only leaves (sampled individuals) of the tree that are grouped (pre-ranked) according to the times they were sampled. A figure is given below (Figure 3) where a subset X_1, X_2, X_3 of set X were sampled before another subset X_4, X_5 of X according to two sampling time.

Although there is a ranking function r, we will consider another pre-ranking function ψ such that $\psi(X_1) = \psi(X_2) = \psi(X_3) = 1$ and $\psi(X_4) = \psi(X_5) = 2$. That means, later sampled individuals will have higher pre-ranking value. Now, we denote the number of individuals sampled at the *i*-th time point with n_i. Note that the number of all fully ranked X-trees depends only on n_i, not on the set X or pre-ranking function ψ.

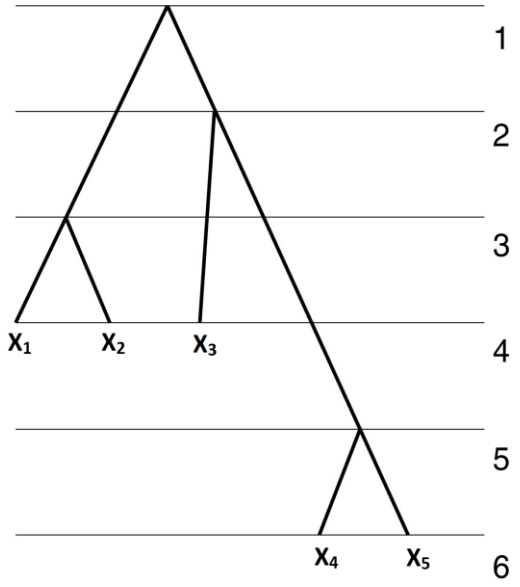

Figure 3: Fully ranked X-tree. $X = X_1, X_2, X_3, X_4, X_5$. The numbers on the right are values of the ranking function [1].

This quantity is expressed by the following function:

$$F(n_{1,\dots,}n_m) = \sum_{i=1}^{n_m} \frac{R(n_m)}{R(i)} F(n_1, \dots, n_{m-1} + 1) \quad (1)$$

And *F(n) = R(n).*

Consider a continuous process of bifurcation in which lineages may bifurcate in time or be cut and labeled (sampled). The process finishes when all lineages are cut producing a tree. The discrete structure of the tree produced by this process is a fully ranked X-tree. It is easy to see that every fully ranked X-tree can be obtained as a result of this process. To count the required number we can count the number of different trees which can be produced by the process if we know that after it finishes there are ni sampled individuals (i.e., cut and labeled lineages) at the i-th time point, i.e., we have the sequence (n_1, ..., n_m)[1].

A third type of tree (Figure 4) can be introduced at this stage, in which sampled individuals may be direct ancestors of later sampled individuals. We call it a tree with sampled ancestors. This type of tree is not usually considered in phylogenetics since the probability of sampling a direct ancestor is often negligible. However, in small populations or when a large portion of the population is sampled, this cannot be ignored. This type of tree is called a fully ranked X-tree with sampled ancestors or FRS X-tree in short.

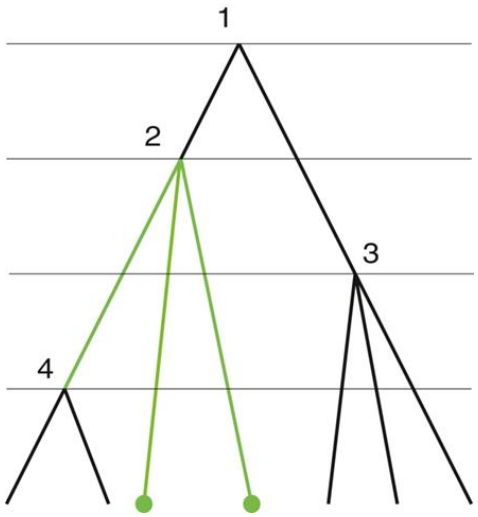

Figure 4: FRS X-tree with the labeled 1-degree root. $X = X_1, X_2, X_3, X_4, X_5$. The numbers on the right are values of the ranking function.

Here, the pre-ranking function remains the same as FRS tree. Let S ($n_1, \dots, n_m$) denote the number of all FRS X-trees that have the same pre-ranking function. Then

$$S(n_1, \dots, n_m) = \sum_{i=1}^{n_m} \sum_{j=0}^{\min(i, n_{m-1})} \binom{i}{j} \binom{n_{m-1}}{j} * \frac{R(n)}{R(i)} S(n_1, \dots, n_{m-1} + i - j) \quad (2)$$

And *S(n) = R(n).*

Now, as we stated earlier, it is quite common to have some prior knowledge about the ancestors of sampled individuals, such as whether there is any monophyletic or crown group existing or not. As we are considering ranked trees, it is obvious that we will have some prior knowledge on relative ages also. Such known information imposes constraints on the space of possible phylogenetic trees representing the evolutionary history of sampled individuals and thus a constraint x-tree is generated. An example is given as Figure 5 with colored edges.

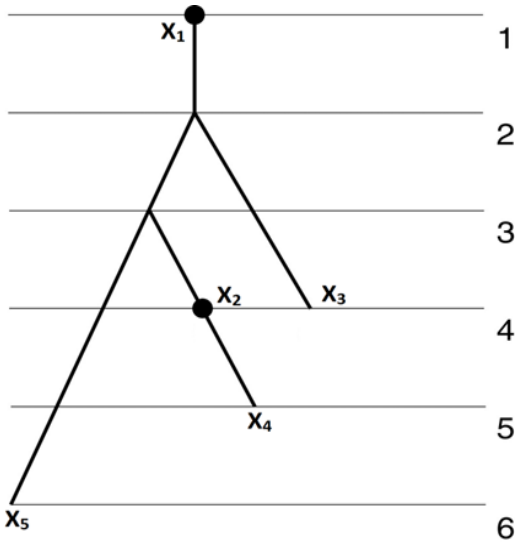

Figure 5: Constraint tree, labels are omitted. Subtree 2 is colored green. It has two child nodes that are leaves, therefore, $n_2 = 2$. The ancestor function for this tree is defined as f (2) = f (3) = 1 and f (4) = 2. A compact notation for this constraint tree is $(n_1, \ldots, n_k, f) = (0, 2, 3, 2, (2, 1), (3, 1), (4, 2))$.

Let $R^r(n_1, \ldots, n_k, f)$ be the number of ranked trees resolving a constraint tree defined by the tuple $(n_1, \ldots, n_k, f)$. Then the following equations hold.

$$R^r(2, \emptyset) = 1 \quad (3)$$

$$R^r(n_1, \ldots, n_{k-1}, 2, f) = \sum_{i \in C} \binom{n_i}{2} * R^r(n_1, \ldots, n_i, 1, \ldots, n_{k-1}, 2, f) + R^r(n_1, \ldots, n_{f(k)} + 1, \ldots, n_{k-1}, f\,|_{2,\ldots,k-1}) \quad (4)$$

$$R^r(n_1, \ldots, n_{k-1}, f) = \sum_{i \in C} \binom{n_i}{2} * R^r(n_2, \ldots, n_{i-1}, \ldots, n_k, f), if\ n_f > 2 \quad (5)$$

[Evaluation of these equations are on [1].]

Here, C is the collection of nodes that have more than 2 children and at least 2 of them are leaves. Therefore, $C = \{i < k | n_i \geq 2$ and $n_i + \beta_i > 2\}$, where $n_i + \beta_i$ = number of children of node i and $\beta_i = |\{b | f(b) = i\}|$.

We will calculate $R^n(n_1, n_2, \ldots, n_k, f)$ for corresponding constraint tree. In order to find it at each step S, we will calculate the numbers of $R^n(X_1, X_{2,\ldots,}X_t, f|_{(t-1)}$ with

$$\sum X_i = s$$
$$i \leq t$$
$$t - 1 = 2, \ldots, t \;\&\; t \leq k \quad (6)$$

Actually we do not have to calculate all such numbers. To determine which numbers are required, we define two upper triangular matrices m and M of size k*k.

Suppose, we draw a horizontal line which is strictly below the line that passes through node j and strictly above the line that passes through node j+1 (or all the leaves if j=k), then

m_{ij} = The minimal possible number of inter- sections of this line with brunches of subtree i,
M_{ij} = The maximal possible number of intersections of this line with brunches of subtree i.

Let $a_{i,j} = |x_j | f(x) = i|$ for ij. So $a_{i,j}$ is the number of children of node i with ranks at most j. Then

$$M_{i,j} = n_i + \alpha_i - a_{i,j}$$
$$m_{i,j} = \begin{cases} 2\ if\ a_{i,j} = 0, \\ 1\ if\ a_{i,j} > 0\ and\ M_{i,j} > 0 \\ 0\ otherwise \end{cases}$$

This leads to the following algorithm to count resolutions. At each step s_n, we construct a set S_s. A unique element of S_n is $R^r(n_1, \ldots, n_k, f)$ and calculating elements of S_s only requires elements of S_{s1}.

Algorithm: Calculating the number of resolutions of a constraint tree

$S_2 = R^r(2, \Phi)$
for $s = 3 \rightarrow (n - 1)$ **do**
  while there is a new element $R^r(x_1, \ldots, x_t, f\,|_{t-1})$ in the set S_{s1} **do**
    if $t < k$ and eligible $(x_1,\ldots,x_{f(t+1)},1,\ldots,x_t,2,f|_{t1})$ **then**
      Calculate $R^r(x_1, \ldots, x_{f(t+1)} - 1, \ldots, x_t, 2, f\,|\,t_1)$ and add it to S_s
    end if
    for $i = 1 \rightarrow t$ **do**
      if eligible$(x_1, \ldots, x_i + 1, \ldots, x_t, f\,|\,_{t1})$ **then**

$R^r(x_1, \ldots, x_i + 1, \ldots, x_t, f \mid_{t1})$
and add it to S_s
end if
end for
end while
end for

From figure 6, the last interior node of a resolving tree (that is, the interior node with the highest rank or the furthest node from the root) is either a parent to leaves in subtree 1 or leaves in subtree 2. Suppose it is the first case (Figure 6, centre).

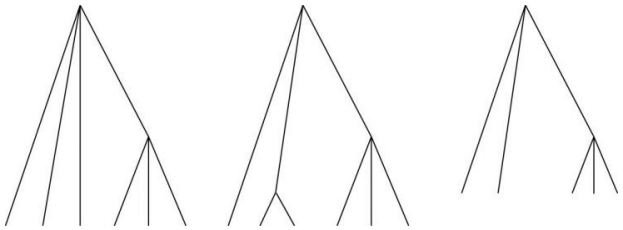

Figure 6: Recursive approach. The last interior node in a resolving tree locates in subtree 2.

Since leaves have distinct labels from X, there are $\binom{3}{2}$ ways to chose two leaves that are children of that last node. We can partition all the resolving trees for which the last node is in subtree 1 in $\binom{3}{2}$ groups. The number of trees in each group is the number of trees that resolve a constraint tree defined by (2, 3, (2, 1)) and shown on the right of Figure 5. A similar argument holds if the last node in a resolving tree is a parent to nodes from subtree 2.
Generalizing the results of constraint X-tree to fully ranked trees, we can also construct fully ranked constraint X-tree or FR Constraint X-tree. However, our study covers a simpler constraint X-tree with three constraints as stated in our introduction section.

III. RESULTS AND DISCUSSION

In phylogenetic analysis, it is common to have some limited information about the ancestors of sampled individuals. This information creates constraint tree. We can generate the tree in more generalized way. Here we have used three constraints like monophyletic group, crown group and relative ages of fossils. This known information imposes constraints on the space of possible phylogenetic trees representing the evolutionary history of sampled individuals and make tree counting more space and time reducing. But on a group of sampled individuals, the number of generated phylogenetic trees satisfying the constraints is still a big issue.

From equation (3), if a constraint tree has 2 leaves then, it is unique and is resolution of itself. We consider the tree with more than 2 leaves in each subtree. However, the last interior node of a resolving tree cannot be in subtree i for $i < k$ if there is not enough leaves in this subtree. This can happen either if there are less than 2 leaves in subtree i or if there are 2 leaves in subtree i and node i has only these two leaves as its children. Both cases imply that any parent to leaves of sub- tree i in a resolving tree have a lesser rank than the rank of node k. Equation (4) explains this is why we need to sum only over the elements of the set C.

If the last node in a constraint tree has only two children, i.e., there are 2 leaves in subtree k, then there is one more group of resolving trees, the group that consists of resolving trees that have this node as the last node according to equation (5).

We generalise these with fully ranked tree and get a multifrucated tree in replace of binary tree. An example shown below on figure 7.

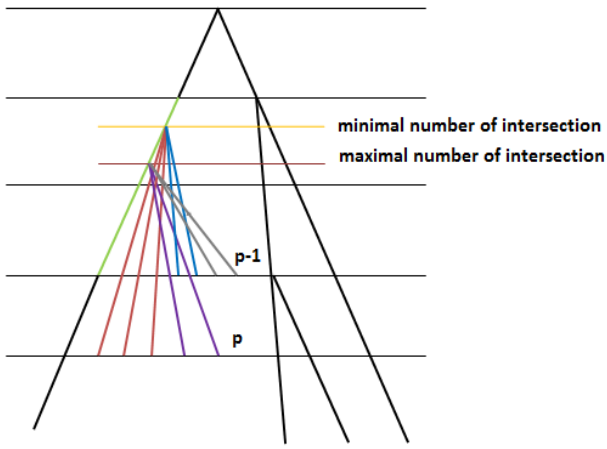

Figure 7: Matrices m and M. Two trees that resolve a constraint tree from Figure 5. The yellow lines correspond to the minimal number of intersections and the red line, to maximal. Location of a new node between the (p − 1)-th and p-th time points. A new node can be placed in subtree between the (p − 1)-th and p-th time points if the number of green branches is greater or equal to 2 and the number of all branches in subtree is greater than 2.

When comparing the calculated numbers of phylogenetic trees (tree counting) for different tree shapes and sizes of leaf sets we recognized that the result actually depends on the constraint of the tree.

IV. CONCLUSION

For contemporaneous sampling the complexity of tree counting algorithm is $O(n)$ and for k constraints $O(n^k)$, which is reasonably fast, particularly when the number of sampling points is small. For Serial sampling, it is $O(mn)$ with no sampled and $O(m^2n^k)$ with k constraints, which is possible if k is small, where n is the sample size, k is the number of constraints, and m is the number of sampling time-points. If we get fossil data serially sampled our proposed tree will be more promising because of tree counting issue.

V. REFERENCES

[1] David Welch Alexandra Gavryushkina1 and Alexei J Drummond. "Recursive algorithms for phylogenetic tree counting". In: *Algorithms for Molecular Biology* 26.8 (2013), pp. 21–36.

[2] Animal Evolution and Diversity. url: http://www.shmoop.com/animal-evolution-diversity/animal-family-tree.html.

[3] Murtagh F. \Counting dendograms: a survey." In: Discrete Appl Math 7 (1984), pp. 191-199.

[4] Murtagh F. \Counting dendograms: a survey." In: Discrete Appl Math 7 (1984), pp. 172-184.

[5] Felsenstein. \The number of evolutionary trees". In: Syst Zool (1978), pp. 27-33.

[6] Heled and Drummod. \Estimating the basic reproductive number from viral sequence data." In: Swiss HIV Cohort 29.11 (2012), pp. 347-357.

[7] Paul Sathio Iwan Sunito. Crown Group Holdings. url: https://en.wikipedia.org/wiki/Crown_Group_Holdings.

[8] Steel M Semple C. \Phylogenetics". In: New York: Oxford University Press (2003).

[9] Stadler T. \Sampling-through-time in birth-death trees." In: J Theor Biol 267(3) (2010), pp. 396-404.

[10] Rannala B Yang Z. \Bayesian phylogenetic inference using DNA sequences: a Markov chain Monte Carlo method." In: Mol Biol Evol 14.7 (1997), pp. 717-724.