Species trees forcing the parsimony to fail modelling evolution process

Vikenty Mikheev
Math. Dept., Kansas State University
Manhattan KS 66506,USA
vikentym@ksu.edu

Serge E. Miheev
Appl.math Dept.
St.Petersburg State University
St.Petersburg 198504, Russia
him2@mail.ru

Abstract

Therefore, Parsimony may be applied only when the described lengths of edges cannot be met in the tree.

1 Introduction

Constructing evolutionary species trees is one of the most interesting problems in biology. It means finding the relations and mutual ancestors of existing and extinct species and also the time of formation of new species. Paleontology itself gives very poor information of species trees structure and time lengths of their edges. More precisely species trees can be built by analysis of genomes of species. One considers the set of species $\{a_i\}_1^N$ and their gene groups $\{G^j\}_1^K$, where $G^j = \{A_i^j\}_{i=1}^N$ is a set of some functionally relative to each other genes. For example, one group can be responsible for hemoglobin production, another one can define the eye color and so on. Let the gene A_i^j be discovered in the species a_i. Then in each j-th functional group one can establish the relations between the genes in the form of unrooted tree, where the leaves are the set of elements of j-th group. The structure of such trees for different groups can coincide (i.e. be topologically identical) or do not coincide. The number of these coincidences defines the frequency of the corresponding gene tree. These frequencies are the base of pasimony method to constuct evolutionary trees which sometimes gives wrong results.

If it is known that a method being applied to some type of problems may fail, why would anyone still use it in this area? Well, in phylogenetics most methods give probabilistic answers. Therefore, getting sometimes wrong answers doesn't necessarily imply that method is bad. To make a final conclusion about the quality of the method one could estimate how often the wrong results appear. Then one would compare the obtained frequency with the frequencies of other methods' failures. Having this information on the table, a researcher can decide if the method is acceptable. However, we did better than this. We have found the set of all combinations of parameters of 5-taxon species tree when Parsimony is guaranteed to fail. Why tree with 5 taxa? It is known that Parsimony always gives right answers on k-taxon species tree for $k = 3, 4$. So, considering $k = 5$ is quite logical from computational point of view. Also the smaller k when things go bad, the louder the warning.

The phylogenetics society has intuitive tendency to use Parsimony less and less. Nevertheless, many biologists still do it because of simplicity of the method. They should not be judged for that since simplicity is a strong argument. The results of this paper will show them the danger of Parsimony. However, forewarned is forearmed. If a researcher is sure that their resulting species tree doesn't have the combination of parameters we showed to be bad, they can safely use fast and simple Parsimony on 5-taxon trees.

For a specific rooted species tree with the known time lengths of the edges, using Coalescence method one can obtain the probabilities of gene trees. That allows to find analytically the rooted species tree and region of

In: Marco Schaerf, Massimo Mecella, Drozdova Viktoria Igorevna, Kalmykov Igor Anatolievich (eds.): Proceedings of REMS 2018 – Russian Federation & Europe Multidisciplinary Symposium on Computer Science and ICT, Stavropol – Dombay, Russia, 15–20 October 2018, published at http://ceur-ws.org

lengths of its edges, when parsimony fails.

2 Preliminaries

We consider an evolution tree T (here and after we mean binary tree) of 5 species a, b, c, d and e with some parameters T_1, T_2, T_3 – time in coalescence units between the branching points (see an example in fig. 1).

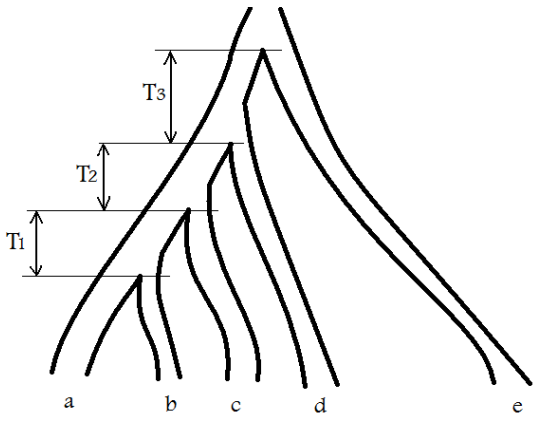
Based on the Coalescent model [Rosenberg], [SemSte], [Wakeley], [Baum] the program COAL [DegnanSalter, WangDegnan] yields the probabilities of all 15 possible unrooted gene trees for 5 genes A, B, C, D, E, such that A is discovered in species a, B is discovered in the species b and so on.

All these species have related genes A, B, C, D, E, respectively. The genes originated from each other or from mutual ancestor. The branching in gene and species trees may not coincide. That creates a problem of inferring species trees from gene trees.

Also, for 5 species there are 15 different unrooted species trees or 105 rooted ones. In each species tree, one can fit any of 15 different gene trees. However, the amount of mutations needed for this fitting will differ generally from one gene tree to another. So, for each gene tree from these 15 and each species tree from the same 15 species trees one can correspond some non-negative integer number of mutations (parsimony score). These can be written in a 15×15 matrix M, where rows correspond to species trees and columns correspond to gene trees.

If one knows the frequencies of different gene trees $P = (p_1, ..., p_{15})^T$, then the mathematical expectation of the number of mutations for each of 15 possible species trees can be calculated by multiplying the matrix M by the vector-column P.

One can assume that the most probable species tree for the given sample of gene trees (or 15 gene trees with assigned probabilities) corresponds to the minimal expectation of mutations. This is the idea of Parsimony method [AllmanRhodes], its application to the problem of inferring species trees from gene trees is called Matrix Representation with Parsimony (MRP) [WangDegnan].



Here is the main question: Does the species tree with the minimal expectation from $M * P$, where P is the vector of probabilities obtained, for example, from the Coalescent model [DegnanSalter] with given T_1, T_2, T_3 for the tree T, present the unrooted version of the original rooted tree T?

Note that we compare rooted species tree with unrooted species tree. It is because the Coalescent method works with rooted trees while parsimony gives only unrooted ones.

It appeared in our work that on the sample of gene trees obtained from a caterpillar species tree with some parameters T_1, T_2, T_3 by coalescence, the parsimony method gives an incorrect species tree.

Figure 1: A 5-taxon species tree with caterpillar topology.

3 Numerical Experiments. Performance of Unrooted MRP for 5-Taxon Species Tree Iinference.

Any 5 genes can be joined in one unrooted tree in 15 ways as given in the following list:

$$\begin{array}{lll} \tau_1\colon ((B,C),A,(D,E)), & \tau_2\colon ((C,D),A,(B,E)), & \tau_3\colon ((C,E),A,(B,D)), \\ \tau_4\colon ((A,E),B,(C,D)), & \tau_5\colon ((A,D),B,(C,E)), & \tau_6\colon ((A,C),B,(D,E)), \\ \tau_7\colon ((A,B),C,(D,E)), & \tau_8\colon ((A,D),C,(B,E)), & \tau_9\colon ((A,E),C,(B,D)), \\ \tau_{10}\colon ((A,B),D,(C,E)), & \tau_{11}\colon ((A,C),D,(B,E)), & \tau_{12}\colon ((A,E),D,(B,C)), \\ \tau_{13}\colon ((A,B),E,(C,D)), & \tau_{14}\colon ((A,C),E,(B,D)), & \tau_{15}\colon ((A,D),E,(B,C)). \end{array}$$

Each unrooted tree may be transformed into a rooted tree by introducing a root to an edge. As a result, we have 7 rooted versions for each unrooted tree.

Step 1. Compute parsimony scores by Fitch-Hartigan [Hartigan] (Table 1).

Thus, if $\vec{N} = (N_1, N_2, ..., N_{15})^{\mathrm{T}}$ is the vector of counts of 15 topological trees in the input, M is the matrix of entries in Table 1 and vector-column $\vec{S} = (pars(\sigma_1),\ pars(\sigma_2), ..., pars(\sigma_{15}))^{\mathrm{T}}$ then $\vec{S} = M\vec{N}$. Here $(pars(\sigma_i)$ is parsimony score of species tree on the collection of gene trees $\tau_1, ..., \tau_{15}$.

Step 2. Pick the smallest entry or entries in $\vec{S}$ to determine the most parsimonious tree(s).

To study the 5-taxon case further we need to use Coalescent Theory. The coalescent model, introduced by Kingman in [Kingmn], describes the coalescence of lineages as we move backwards in time within a single species (Note that in biology the understanding of the word 'species' may vary. Here we use this word in the same meaning as 'population'). By "gluing" together such species or populations to form a tree, one gets the Multi-species Coalescent Model, which describes the production of gene trees within species trees.

Table 1: The parsimony scores $pars_{\tau_j}(\sigma_i) \equiv m_{ij}$ for all 15 possible output trees σ with respect to the matrix representation of all 15 possible input trees τ.

	τ_1	τ_2	τ_3	τ_4	τ_5	τ_6	τ_7	τ_8	τ_9	τ_{10}	τ_{11}	τ_{12}	τ_{13}	τ_{14}	τ_{15}
σ_1	2	4	4	4	4	3	3	4	4	4	4	3	4	4	3
σ_2	4	2	4	3	4	4	4	3	4	4	3	4	3	4	4
σ_3	4	4	2	4	3	4	4	4	3	3	4	4	4	3	4
σ_4	4	3	4	2	4	4	4	4	3	4	4	3	3	4	4
σ_5	4	4	3	4	2	4	4	3	4	3	4	4	4	4	3
σ_6	3	4	4	4	4	2	3	4	4	4	3	4	4	3	4
σ_7	3	4	4	4	4	3	2	4	4	3	4	4	3	4	4
σ_8	4	3	4	4	3	4	4	2	4	4	3	4	4	4	3
σ_9	4	4	3	3	4	4	4	4	2	4	4	3	4	3	4
σ_{10}	4	4	3	4	3	4	3	4	4	2	4	4	3	4	4
σ_{11}	4	3	4	4	4	3	4	3	4	4	2	4	4	3	4
σ_{12}	3	4	4	3	4	4	4	4	3	4	4	2	4	4	3
σ_{13}	4	3	4	3	4	4	3	4	4	3	4	4	2	4	4
σ_{14}	4	4	3	4	4	3	4	4	3	4	3	4	4	2	4
σ_{15}	3	4	4	4	3	4	4	3	4	4	4	3	4	4	2

Definition 1. Let $g_{ij}(T)$ denote the probability that i lineages (genes) since time 0 have coalesced to exactly j lineages at time T under the coalescent model.

General formulas for the $g_{ij}(T)$ were derived in [Tavaré]:

$$g_{ij}(T) = \sum_{k=j}^{i} e^{-k(k-1)T/2} \frac{(2k-1)(-1)^{k-j} j_{(k-1)} i_{[k]}}{j!(k-j)! i_{(k)}}, \tag{1}$$

where $a_{(k)} = a(a+1)\cdots(a+k-1)$ for $k \geq 1$ with $a_{(0)} = 1$ (the partial permutation); and $a_{[k]} = a(a-1)\cdots(a-k+1)$ for $k \geq 1$ with $a_{[0]} = 1$.

Some of these formulas for small indexes are

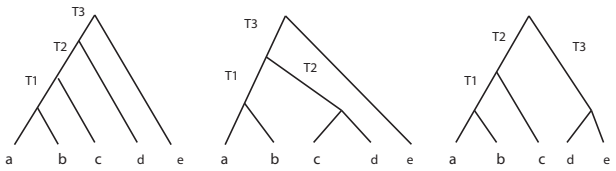
$g_{11}(T) = 1$, $g_{21}(T) = 1 - e^{-T}$, $g_{22}(T) = e^{-T}$,

$g_{31}(T) = 1 - (3/2)e^{-T} + (1/2)e^{-3T}$, $g_{32}(T) = (3/2)e^{-T} - (3/2)e^{-3T}$, $g_{33}(T) = e^{-3T}$,

$g_{41}(T) = 1 - (9/5)e^{-T} + e^{-3T} - (1/5)e^{-6T}$, $g_{42}(T) = (9/5)e^{-T} - 3e^{-3T} + (6/5)e^{-6T}$,

$g_{43}(T) = 2e^{-3T} - 2e^{-6T}$, $g_{44}(T) = e^{-6T}$

Let 3 rooted species tree be $\Sigma_1 := ((((a,b):T_1,c):T_2,d):T_3,e)$, $\Sigma_2 := (((a,b):T_1,(c,d):T_2):T_3,e)$ and $\Sigma_3 := (((a,b):T_1,c):T_2,(d,e):T_3)$. They are rooted versions of σ_7, σ_{13} and σ_7 again, respectively (see Fig. 2).



Up to taxon names, these three are the only possible species trees. They are usually referred to as the *caterpillar* (Σ_1), *pseudo-caterpillar* (Σ_2) and *pseudo-balanced* (Σ_3) species trees.

Figure 2: Three rooted 5-taxon species trees Σ_1, Σ_2 and Σ_3.

4 An experiment with caterpillar

Using the program COAL [DegnanSalter, WangDegnan] to get probabilities of rooted gene trees and the formulas (1) for g_{ij}, we calculate the probabilities $p_i = p(\tau_i|\Sigma_1)$ for $i = \overline{1,15}$, which are listed in Table 2 ($X := e^{-T_1},\ Y := e^{-T_2},\ Z := e^{-T_3}$,) after simplification in Maple 15.

Table 2: The probabilities $p_i = p(\tau_i|\Sigma_1)$ for $i = \overline{1,15}$, where $X = e^{-T_1}$, $Y = e^{-T_2}$, $Z = e^{-T_3}$.

p_1	$X/3 - XY/3 + XY^3/18 + XY^3Z^6/90$
p_2	$XY^3/18 + XY^3Z^6/90$
p_3	$XY^3/18 + XY^3Z^6/90$
p_4	$XY^3/18 + XY^3Z^6/90$
p_5	$XY^3/18 + XY^3Z^6/90$
p_6	$X/3 - XY/3 + XY^3/18 + XY^3Z^6/90$
p_7	$1 - 2X/3 - 2Y/3 + XY/3 + XY^3/18 + XY^3Z^6/90$
p_8	$XY^3/18 + XY^3Z^6/90$
p_9	$XY^3/18 + XY^3Z^6/90$
p_{10}	$Y/3 - XY/6 - XY^3/9 + XY^3Z^6/90$
p_{11}	$XY/6 - XY^3/9 + XY^3Z^6/90$
p_{12}	$XY/6 - XY^3/9 + XY^3Z^6/90$
p_{13}	$Y/3 - XY/6 - XY^3/18 - 2XY^3Z^6/45$
p_{14}	$XY/6 - XY^3/18 - 2XY^3Z^6/45$
p_{15}	$XY/6 - XY^3/18 - 2XY^3Z^6/45$

Let vector-column $\mathbf{p}$ be $(p_1, p_2, \cdots, p_{15})^T$. We consider the product $\mathbf{s}^{cat} := M\mathbf{p}$, where each entry is the expectation of parsimony score of a possible output tree for MRP.

We discover that in $\mathbf{s}^{cat}$ some entries are equal. Let's denote them as following

$\alpha^{cat} := s_1^{cat} = s_6^{cat} = 3 - X/3 + 2Y/3 + XY/3 - XY^3/18 - XY^3Z^6/90,$
$\beta^{cat} := s_2^{cat} = s_3^{cat} = s_4^{cat} = s_5^{cat} = 4 - Y/3 - XY^3/18 - XY^3Z^6/90,$
$\gamma^{cat} := s_7^{cat} := 2 + 2X/3 + 2Y/3 + XY/3 - XY^3/18 - XY^3Z^6/90,$
$\delta^{cat} := s_8^{cat} = s_9^{cat} = 4 - XY/3 - XY^3/18 - XY^3Z^6/90,$
$\epsilon^{cat} := s_{10}^{cat} := 3 + 2X/3 - Y/3 + XY/6 + XY^3/9 - XY^3Z^6/90,$
$\zeta^{cat} := s_{11}^{cat} = s_{12}^{cat} = 4 - X/3 - XY/6 + XY^3/9 - XY^3Z^6/90,$
$\eta^{cat} := s_{13}^{cat} := 3 + 2X/3 - Y/3 + XY/6 + XY^3/18 + 2XY^3Z^6/45,$
$\theta^{cat} := s_{14}^{cat} = s_{15}^{cat} = 4 - X/3 - XY/6 + XY^3/18 + 2XY^3Z^6/45.$

The analytical comparison of these values we form in the following

Proposition 1. *For any $X, Y, Z \in (0,1)$, the following inequalities hold:*

$$\gamma^{cat} < \zeta^{cat}, \quad \eta^{cat} < \theta^{cat}, \quad \gamma^{cat} < \alpha^{cat} < \beta^{cat}, \quad \gamma^{cat} < \delta^{cat}, \quad \gamma^{cat} < \epsilon^{cat}.$$

However, the 3D-graphs on Figures 3 show that the expressions η^{cat} and γ^{cat} can not be put in one order for all $X, Y, Z \in (0,1)$.

There is a large region were $\gamma^{cat} < \eta^{cat}$ but, nevertheless, there is also a region where $\eta^{cat} < \gamma^{cat}$. The last defines the parameters T_1, T_2, T_3 where MRP will fail to recover the tree topology of the true species tree producing the gene tree distribution, even when given an arbitrary large sample of gene trees. Figure 3.left shows that provided Y is not too large, regardless of X, Z, MRP will return the correct species tree.

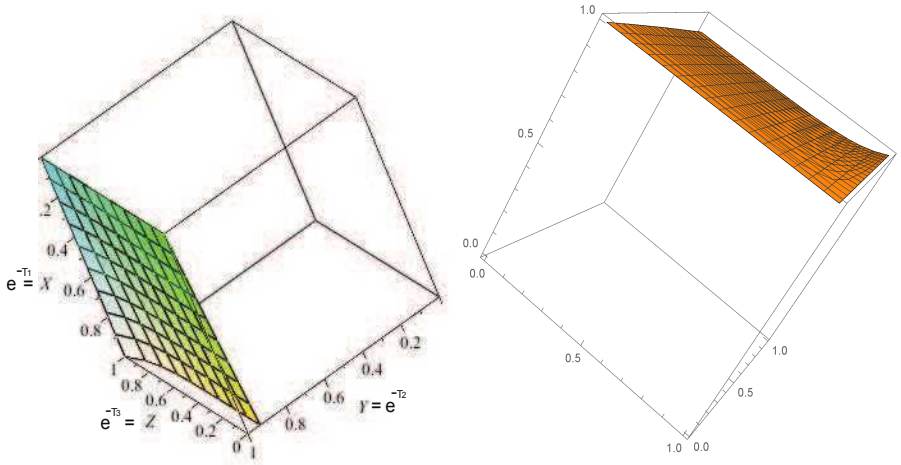


Figure 3: Left: the surface $\eta^{cat}(X,Y,Z) = \gamma^{cat}(X,Y,Z)$. $\eta^{cat} > \gamma^{cat}$ on the large region including all those points, when Y is near 0, while $\eta^{cat} < \gamma^{cat}$ on the small region. Right: Different angle on the same surface $\eta^{cat}(X,Y,Z) = \gamma^{cat}(X,Y,Z)$.

To determine this cutoff for Y, we set $\eta^{cat}(1,Y,0) = \gamma^{cat}(1,Y,0)$ and solve to get $Y = 0.935...$ (the solutions of $(1/3)Y^3 - (7/2)Y + 3 = 0$ are $Y_{1,2,3} \approx 2.670..., -3.605..., 0.935...$).

5 An experiment with pseudo-caterpillar

Let the pseudo-caterpillar species tree be $\Sigma_2 = (((a,b):T_1,(c,d):T_2):T_3,e)$. We use COAL, the formulas for g_{ij} and Maple 15 to calculate the probabilities $p_i^{pc} = p(\tau_i|\Sigma_2)$ for $i = \overline{1,15}$. Their list is shown in Table 3 after simplifications.

Table 3: The probabilities $p_i^{pc} = p(\tau_i|\Sigma_2)$ for $i = \overline{1,15}$, where $X = e^{-T_1}$,.

$p_1^{pc} = p_3^{pc} = p_5^{pc} = p_6^{pc} = p_8^{pc} = p_9^{pc} = p_{11}^{pc} = p_{12}^{pc} = (1/18)XY + XYZ^6/90,$
$p_2^{pc} = p_4^{pc} = p_7^{pc} = p_{10}^{pc} = XYZ^6/90 + Y/3 - (5/18)XY,$
$p_{13}^{pc} = 1 - (2/45)XYZ^6 + (4/9)XY - (2/3)X - (2/3)Y,$
$p_{14}^{pc} = p_{15}^{pc} = -(2/45)XYZ^6 + XY/9.$

In the table 3: $X = e^{-T_1}$, $Y = e^{-T_2}$, $Z = e^{-T_3}$

Now assuming $\mathbf{p}^{pc} = (p_1^{pc}, p_2^{pc}, \cdots, p_{15}^{pc})^T$ we see that the product $\mathbf{s}^{pc} := M\mathbf{p}^{pc}$ is the vector of expected parsimony scores of possible output trees with a pseudo-caterpillar species tree. We discover that in $\mathbf{s}^{pc}$ some entries are equal. Let's denote them as following

$\alpha^{pc} := s_1^{pc} = s_3^{pc} = s_5^{pc} = s_6^{pc} = 4 - XYZ^6/90 - Y/3 - XY/18,$
$\beta^{pc} := s_2^{pc} = s_4^{pc} = 3 - XYZ^6/90 - X/3 + 2Y/3 + 5XY/18,$
$\gamma^{pc} := s_7^{pc} = s_{10}^{pc} = 3 - XYZ^6/90 + 2X/3 - Y/3 + 5XY/18,$
$\delta^{pc} := s_8^{pc} = s_9^{pc} = s_{11}^{pc} = s_{12}^{pc} = 4 - XYZ^6/90 - X/3 - XY/18,$
$\zeta^{pc} := s_{13}^{pc} = 2 + 2XYZ^6/45 + 2X/3 + 2Y/3 + 2XY/9,$
$\eta^{pc} := s_{14}^{pc} = s_{15}^{pc} = 4 + 2XYZ^6/45 - 4XY/9.$

Let's compare these expressions.

Proposition 2. *For any $X, Y, Z \in (0,1)$, the following inequalities hold:*

$$\zeta^{pc} < \alpha^{pc},\ \zeta^{pc} < \beta^{pc},\ \zeta^{pc} < \gamma^{pc},\ \zeta^{pc} < \delta^{pc},\ \zeta^{pc} < \eta^{pc}.$$

This implies that if the true species tree is 5-taxon pseudo-caterpillar, MRP, for a sufficiently large data set, will give with probability 1 the unrooted species tree topology for all $T_1, T_2, T_2 \in (0, 1)$.

6 An experiment with psuedo-balanced

The last tree we need to consider (since all other are just permutations of taxon names) is the pseudo-balanced species tree $\Sigma_3 = (((a, b) : T_1, c) : T_2, (d, e) : T_3)$. The chain of actions is exactly the same that in previous two experiments. We omit it here. The result is Parsimony doesn't fail.

So, if the true species tree is 5-taxon pseudo-balanced Σ_3, MRP, for a sufficiently large data set, will give with probability 1 the correct unrooted species tree topology for all $T_1, T_2, T_2 \in (0, 1)$.

7 Generalization of results.

7.1 Caterpillar Subtree.

Definition 2. There is a rooted tree T with number of taxa equal to $|T| =: n$. Let $T_{cat}(T)$ be a caterpillar subtree of this tree. The number $Cat(T) := \max\limits_{T_{cat}(T) \subset T} |T_{cat}(T)|$ for a particular tree T is called *caterpillar score for the tree* T. The number $cat(n) := \min\limits_{|T|=n} Cat(T)$ is called *caterpillar measure.*

It is clear that $cat(n)$ is an increasing function with respect to n. There are may be a few consecutive numbers n such that $cat(n) = k$ for some given natural k.

Definition 3. Let's call number $r_k := \min\limits_{cat(n)=k} n$ the *revolution number.*

Observe that for the caterpillar lengths $1, 2, 3$ their revolution numbers are $r_1 = 1, r_2 = 2, r_3 = 3$, because these trees are caterpillar themselves. Note, that the third and the second revolution numbers are connected by

$$r_3 = 2r_2 - 1. \tag{2}$$

Theorem 1. *The revolution numbers r_k may be calculated recursively $r_k = 2r_{k-1} - 1$, $k = 4, 5....$*

Proof. Let T be a k-taxon tree. Since T is a binary tree, we can think of T as two subtrees T', T'' glued together only by two edges at the root (see Figure 4). Observe that for any caterpillar subtree of T' one of these T', T'' being transformed properly brings only one edge to the caterpillar.

On the other hand, r_k is an increasing function. So, the first bifurcation in the root will be the worst in the sense of caterpillar score for the tree T when this bifurcation divides the tree into two subtrees T' and T'' such that $|T'| = |T''|$ for even $|T|$, and $\big||T'| - |T''|\big| = 1$ for odd $|T|$. Further, we use mathematical induction.

The base of induction. It is the formula (2)

The assumption of induction. Let $r_3, ..., r_k$ calculated recursively be revolution numbers.

The inductive step. We need to prove that $r_{k+1} = 2r_k - 1$ is the revolution number. Let us take a tree T with r_{k+1} taxa and consider the worst bifurcation in its root. As mentioned above, the worst bifurcation in the root of T forms two subtrees T' and T'' such that $|T'| = r_k - 1,\ |T''| = r_k$.

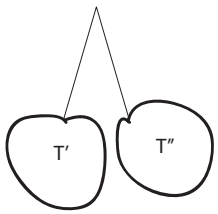


Figure 4:

The induction assumption yields an existence of caterpillar in T'' with a length no less than k. One edge in T' together with the caterpillar in T'' creates the caterpillar tree C with $|C| = k + 1$. So, the revolution number for $k + 1$ is no greater than $2r_k - 1$. If $|T| \in [r_k, 2r_k - 1)$, then the worst bifurcation forms two subtrees T' and T'' such that $|T'|, |T''| < r_k$. Since r_k is a revolution number, there are T' and T'' which have only caterpillars C', C'' and C'' with $|C'|, |C''| < r_k$. Therefore, r_{k+1} is the revolution number.

Theorem 7.1 allows to continue the sequence of the revolution numbers for the caterpillar measures $3, 4, 5, 6, 7, ...$ as $r_k = 3, 5, 9, 17, 33, ...$, respectively. For trees with number of taxa in $[r_k, 2r_k - 1]$, the caterpillar measure is k.

7.2 MRP on trees with the number of taxa greater than 5

Theorem 2. *If a true species rooted tree $\hat{G}$ contains 5-taxon caterpillar subtree, then MRP may fail to obtain the unrooted version of $\hat{G}$ from the set of gene trees generated by Coalescent model from $\hat{G}$.*

Proof. For the number of species greater than 5 and the same number of genes one can make the following construction.

Take a caterpillar tree Γ of 5 species a, b, c, d, e with T_1, T_2, T_3 and root ρ, such that parsimony fails (fig. 3). Take an arbitrary tree G, where every edge is very close to 0 ($T_i \approx 0,\ \ i > 4$). Connect Γ to G through its root ρ and the edge ϵ with the length T_4. Make T_4 big enough so the genes A, B, C, D, E coalesce in ϵ if they didn't in Γ. No matter what is on the upper end of ϵ, root of entire tree or inner node created by ϵ on some edge of G .

The numeration of n possible gene trees we do in the following way: First 15 trees will have the same subtree G and different topology or permutation of A, B, C, D, E. Other $n - 16$ trees can be numerated in any order, and we set their probabilities $p_{16}, ..., p_n$ equal to zero, since $T_i,\ \ i \in \{5, ..., n\}$, can be taken infinitively small. Therefore, the set of gene trees is numerated and the probabilities of them are presented by vector-column $p = (p_1, ..., p_{15}, p_{16}, ..., p_n)^T$, where all the entries below 15-th equal zero and the first 15 are the same that obtained ones for 5-taxon experiment. The matrix $M_{n\times n}$ has dimension $n \times n$, but only submatrix $M_{n\times 15}$ does participate in calculation of expectations of gene mutations due to $p_{16}, ..., p_n = 0$.

Moreover, the parsimony incorrect choice may be shown on submatrix $M_{15\times 15}$ in upper corner. Observe that the elements of $M_{15\times 15}$ are the sums of the elements of M obtained earlier in performance of MRP for 5-taxon trees 1 and some constant number generated by the constant subtree G with coalesced gene $A+B+C+D+E$. This means that each of s_1^{cat}, s_2^{cat}, ... , s_{15}^{cat} from Section 4 must be increased by the constant value $\ V\sum_1^{15} p_i$, to be a new mathematical expectation for the new big tree. So, the minimum among the first 15 rows must be achieved in the same index. Therefore, being wrong for 5-taxon caterpillar species tree the parsimony becomes wrong for the constructed tree $\hat{G}$ as well.

Corollary 1. *MRP on a set of gene trees with 5 taxa or more may yield wrong result. If one applies MRP on set of gene trees with 9 taxa or more, MRP may fail even more probably, since 9-taxon species tree always has a caterpillar subtree, which may have unfortunate lengths of inner edges from the small region in Figure 3*

We have established what these unfortunate lengths are. But how to find these caterpillar subtrees? One may use the following

Theorem 3. *If a tree has three consecutive inner edges not contaning the root betwen them but perhaps one of these edges ending on it then the tree has a 5-taxon caterpillar subtree, which contains these three inner edges.*

Proof. At Figure 5 we can see three consecutive edges denoted 1,2 and 3 between nodes a,b, c and d, respectively and the third edge is closest to the root. Since these edges are inner, all the nodes must be bifurcation points and so each of b, c, d has one more edge running towards a taxon or a clade opposite to the root and a has two more such edges. Let d be the root then simply contracting each of the mentioned clades to one of its taxa we get a 5-taxon caterpillar subtree. If d is not the root, throw away one of the edges at d to make d the root of the 5-taxon caterpillar subtree.

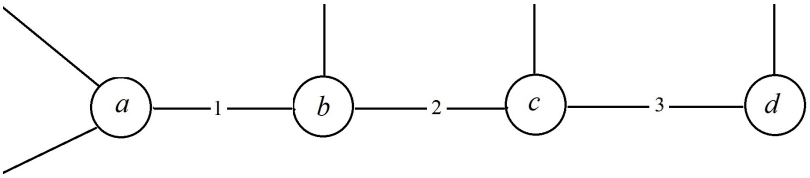


Figure 5: Three consecutive inner edges in some tree.

Corollary 2.*It is enough to know that in a species tree with amount of taxa 5 or more there are three consecutive inner edges not going through the root but perhaps ending on it with lengths T_1, T_2 and T_3 from the small region of cube in Figure 3 to conclude that Parsimony is guaranteed to fail on this tree.*

8 Conclusions and Future work

The fact that Parsimony may fail is not new. However, here we proved that no mater of what topology the true 9-taxon and greater species tree is the only condition to fail Parsimony is to have in this tree three consecutive

inner edges not going through the root but perhaps ending on it with lengths $T1, T2, T3$ (which are times in coalescence units between the branching points) of some proportions. Obviously, the probability to meet these lengths is growing in general with the size of species tree. So, if one wants to safely use MRP on a set of n-taxon gene trees, it is need to know somehow that the resulting n-taxon species tree cannot have any of "bad" topologies and edge lengths from "bad" regions. This paper makes it possible for $n \leq 5$. Also, one may apply MRP on a set of 5-taxon gene trees and if the result is the caterpillar tree or topology Σ_3 from Fig. 2, it is true.

One may consider 6-taxon species trees the way we did in this paper and check the existence of 6-taxon topology which forces Parsimony to fail when lengths of inner edges have some proportions. Then prove, perhaps following our ideas, that every tree with some number of taxa greater than 6 has this 6-taxon topology subgraph. Then do the same for 7 taxa and so on.

If one studies all "bad" parameters' regions of all "bad" topologies for all trees with the amount of taxa less or equal some n, it becomes theoretically possible to check either MRP can be applied for a set of gene trees of n taxa (if, of course, the researcher knows enough information about possible results). However, taking into account the factorial growth of the amount of binary trees with respect to the amount of taxa, the problem to find "safe zone" for MRP becomes extremely hard. Unfortunately, we don't see any way around besides doing that scheme for each $k < n$. So, someone has to be very motivated to use MRP to go through with the research. Nowadays, there are good coalescence based methods, for example, [YufengWu], [RanYan] and [EmmsKelly]

It could be interesting to study stability questions of the coalescent model with uncertainty in data applying the thoughts from [Zubov].

References

[AllmanRhodes] E. S. Allman and J. A. Rhodes. *Lecture Notes: The Mathematics of Phylogenetics*. University of Alaska Fairbanks, 2009.

[Baum] B. R. Baum. Combining trees as a way of combining data sets for phylogenetic inference, and the desirability of combining gene trees. *Taxon*, 41:3 – 10, 1992.

[DegnanSalter] J. H. Degnan and L. A. Salter. Gene tree distributions under the coalescent process. *Evolution*, 59(1):24 – 37, 2005.

[Hartigan] J. A. Hartigan. Minimum mutation fits to a given tree. *Biometrics*, 29:53 – 65, 1973.

[Kingmn] J. F. C. Kingman. The coalescent. *Stoch. Process. Appl.*, 13:235 – 248, 1982.

[Rosenberg] N. A. Rosenberg. The probability of topological concordance of gene trees and species trees. *Theor. Pop. Biol.*, 61:225 – 247, 2002.

[SemSte] Charles Semple and Mike Steel. *Phylogenetics*, volume 24 of *Oxford Lecture Series in Mathematics and its Applications*. Oxford Un. Press, Oxford, 2003.

[Tavaré] S. Tavaré. Line-of-descent and genealogical processes, and their applications in population genetics models. *Theoret. Population Biol.*, 26(2):119–164, 1984.

[Wakeley] J. Wakeley. *Coalescent Theory*. Roberts & Company, Greenwood Village, CO, 2008.

[WangDegnan] Yuancheng Wang and James H. Degnan. Performance of matrix representation with parsimony for inferring species from gene trees. *Stat. Appl. Genet. Mol. Biol.*, 10:Art. 21, 41, 2011.

[Zubov] I. V. Zubov and A. V. Zubov. The stability of motion of dynamic systems. *Doklady Mathematics*, 79(1):112 – 113, 2009.

[YufengWu] Yufeng Wu. A coalescent-based method for population tree inference with haplotypes *Bioinformatics*, 31(5):691 – 698, 2015.

[RanYan] Bruce Rannala and Ziheng Yang. Efficient Bayesian Species Tree Inference under the Multispecies Coalescent *Systematic Biology*, 66(5): 823–842, 2017.

[EmmsKelly] David Emms and Steven Kelly. STAG: Species Tree Inference from All Genes *Biorxiv* https://doi.org/10.1101/267914 Published Feb. 19, 2018.