MASTER'S DEGREE THESIS

Master of Science in Advanced Mathematics and Mathematical Engineering

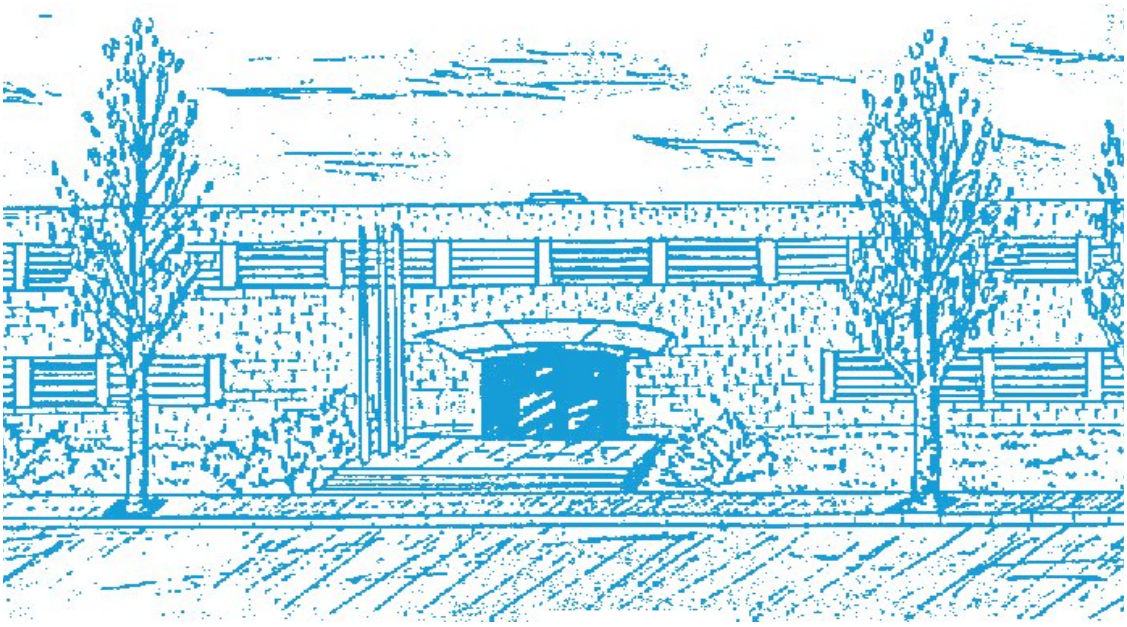
Title: Implementation of a Test for Branch Saturation in a Phylogenetic Tree

Author: Carmen Cinta Viñas Ferrando

Advisor: Prof. Arndt Von Haeseler and Cassius Manuel Pérez de los Cobos Hermosa

Department: Departament de Matemàtica Aplicada I

Academic year: 2021-2022



UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat de Matemàtiques i Estadística

Universitat Politècnica de Catalunya
Facultat de Matemàtiques i Estadística

Master in Advanced Mathematics
and Mathematical Engineering

Master's Thesis

Implementation of a Test for Branch Saturation in a Phylogenetic Tree

Carmen Cinta Viñas Ferrando

Supervised by Prof. Arndt Von Haeseler
and Cassius Manuel Pérez de los Cobos Hermosa

September, 2022

First of all, I would like to thank Cassius for guiding me through this project and advising me whenever it was needed. I would also like to thank Prof. Arndt von Haeseler for welcoming me at the CIBIV and having provided advise and expertise, and Prof. Marta Casanellas for her support.

ABSTRACT

In a phylogenetic tree, branch saturation is the occurrence of too many mutations along a branch as to provide information about the branch location and length. Since branch saturation leads to unreliable reconstructed trees, testing all branches for saturation is a necessary step in any accurate reconstruction protocol. The core of this work is the software implementation of the asymptotic test for branch saturation. After presenting three examples using different trees, the concept of saturation and its causes are analysed.

Keywords: Phylogenetic Tree, Maximum Likelihood, Branch Saturation, Reconstruction

Contents

1 Introduction

All living organisms share common genetic material. Hence, species are related to each other by evolutionary relationships, which can be represented using a phylogenetic tree. The goal of phylogenetics is to reconstruct what cannot be observed, that is, the evolutionary process and the ancestral sequences. More precisely, phylogenetic reconstruction aims to identify the regions of the DNA of different species which contain analogous information, describe the evolutionary relationships between such species in terms of relative recency of common ancestry and recover the evolutionary distance between them.

There are several methods of phylogenetic reconstruction. In this project, we focus on the maximum likelihood method, which computes the probability that a data set fits a tree and model of sequence evolution.

Since ancestral sequences cannot be observed, the reliability of their reconstruction depends on the amount of information carried and preserved from the leaves. For this purpose, Manuel and von Haeseler [2022] define the memory as the amount of identification of the ancestral sequence, and the coherence of a given branch as the amount of dependence between the two nodes separated by the branch.

These two measures are employed by Manuel and von Haeseler [2022] to construct a test to detect branch saturation. Intuitively, saturation is the occurrence of too many mutations in an alignment as to provide information about its evolutionary history.

The aim of this project is to understand phylogenetic reconstruction using maximum likelihood, as well as the new concepts describing phylogenetic information, namely memory, coherence and saturation. In addition, we implement the saturation test in an efficient and systematic way. Then we execute our implementation in different trees to have a better comprehension of the concept of branch saturation.

The memoir is divided in six sections. After the introduction, Section 2 is dedicated to explain the basic phylogenetic concepts, as well as the likelihood computation in a phylogenetic tree. In Section 3 we explain the new measures of phylogenetic information and the asymptotic test for the detection of branch saturation introduced by Manuel and von Haeseler [2022]. Sections 4 and 5 are devoted to the explanation of the methods used and the results obtained, respectively. Finally, in Section 6 we summarise the conclusions of the project.

The code of the implementation of the test has been written in Python, and IQ-TREE [B.Q. et al., 2020] is used for the reconstruction.

2 Preliminaries

2.1 Phylogenetic Trees

A **phylogenetic tree** $\mathcal{T}$ is a connected graph with no cycles where each node represents a species or taxon. In a phylogenetic tree, the external nodes are called **leaves** and depict the currently existing species, while the internal nodes represent (possibly extinct) ancestral species. The **edges** of the tree represent the evolutionary processes connecting the nodes.

A topology on a phylogenetic tree is described as the shape of the tree in addition to a particular labelling of the leaves. In other words, it is a fixation of the branching structure of the tree, thus establishing patterns of relatedness among taxa. If two phylogenetic trees display the same topology and root, then they can be said to share the same biological interpretation.

While the topology of the tree represents the speciation events occurred along the evolutionary history, the length of an edge represents the number of mutations that occur between the two species at the ends of such edge.

A tree is called a rooted tree if one vertex has been labelled as root, which represents the common ancestor to all entities in the tree, and the edges are oriented away from it. Otherwise, we say the tree is unrooted.

The number of branches connected to a node is called the **degree** of the node. Leaves have degree 1. We say that a tree is binary if all internal nodes have degree 3.

Given a tree with m leaves and the set of sequences $(\boldsymbol{l^1}, ..., \boldsymbol{l^m})$ at the leaves, each i'th site $\partial \in \mathbb{A}^m$, where $\mathbb{A}$ is the state space of the entries at the leaves, is called a **pattern**. That is, the set of nucleotides located in the i'th position of every leaf forms a pattern, as represented in Figure 1. An alignment has as many patterns as the number of nucleotides that are in a leaf sequence, and each pattern has a length equal to the number of leaves.

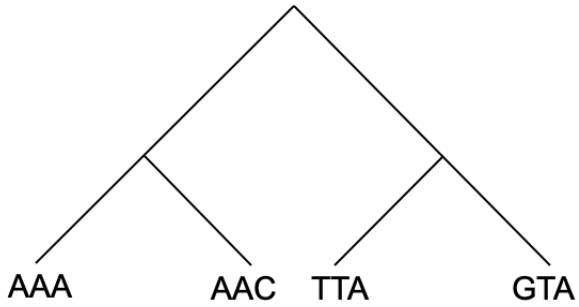


Figure 1: Diagram of a phylogenetic tree with four leaves, each having a sequence of three nucleotides. The concatenation of the i'th nucleotide from each leaf forms a pattern. In this example four patterns are observed: $AATG$, $AATT$ and $ACAA$.

2.2 Evolutionary Models

An evolutionary model describes the process through which a sequence of nucleotides evolves into one another as a series of random mutational events. Evolutionary models consider that processes at different tree edges are independent and nucleotides in a DNA sequence do not depend on each other. In other words, evolutionary models assume that evolution in different sites is independent and evolution in different lineages is independent.

In continuous Markov models of evolution, each site of a sequence mutates independently according to a probability transition matrix $P(t) = e^{Qt}$, with Q being the rate matrix.

Definition 2.1. A **rate matrix** $Q = (q_{ij})$ is a square matrix where the entries q_{ij} describe the instantaneous rate at which the state i mutates to state j. Each row in the rate matrix sums to 0 and the diagonal elements are defined as $q_{ii} = -\sum_{i \neq j} q_{ij}$. A rate matrix Q satisfies the following conditions:

(1) $q_{ii} \leq 0$ for all i
(2) $q_{ij} \geq 0$ for any $i \neq j$
(3) $\sum_j q_{ij} = 0$ for all i

Definition 2.2. A **probability transition matrix** $P(t) = (p_{ij}(t)) = e^{Qt}$ is a square matrix where the entries $p_{ij}(t)$ describe the probability of state i mutating to state j in time t. The sum of the elements of each row in the probability transition matrix must be 1.

The state space of each nucleotide is $\mathbb{A} = \{A, C, G, T\}$, each character corresponding to the one of the four nucleotides that constitute the DNA. The nucleotide distribution at time $t \geq 0$ is the column vector $\boldsymbol{\pi}(t) = (\pi_A(t), \pi_C(t), \pi_G(t), \pi_T(t))$ and the initial nucleotide distribution is $\boldsymbol{\pi}(0)$. At any time $t > 0$, the nucleotide distribution satisfies $\boldsymbol{\pi}(t)^T = \boldsymbol{\pi}^T(0)P(t)$.

Definition 2.3. An evolutionary process is **stationary** if the prior state distribution is the unique equilibrium nucleotide distribution $\boldsymbol{\pi}$ that fulfils $\boldsymbol{\pi}^T Q = 0$ and $\boldsymbol{\pi}^T P(t) = \boldsymbol{\pi}$.

Definition 2.4. An evolutionary process is **time-reversible** if $\pi_i q_{ij} = \pi_j q_{ji}$ for all $i \neq j$.

In this project we consider stationary and reversible evolutionary models.

Definition 2.5. The **JC69** (Jukes and Cantor 1969) model is the simplest evolutionary model, which assumes equal base frequencies $\left(\boldsymbol{\pi} = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})\right)$ and has a rate matrix of the form

$$Q = \begin{pmatrix} \cdot & \frac{\mu}{3} & \frac{\mu}{3} & \frac{\mu}{3} \\ \frac{\mu}{3} & \cdot & \frac{\mu}{3} & \frac{\mu}{3} \\ \frac{\mu}{3} & \frac{\mu}{3} & \cdot & \frac{\mu}{3} \\ \frac{\mu}{3} & \frac{\mu}{3} & \frac{\mu}{3} & \cdot \end{pmatrix}, \text{ with } \mu \in [0, 1].$$

Definition 2.6. The **K80** (Kimura 1980) or **K2P** (Kimura two parameter) model assumes equal base frequencies $\left(\boldsymbol{\pi} = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})\right)$. It distinguishes between transitions ($A \leftrightarrow G$ and $T \leftrightarrow C$) and transversions ($A \leftrightarrow T$, $A \leftrightarrow C$, $G \leftrightarrow T$ and $G \leftrightarrow C$) and has a rate matrix of the form

$$Q = \begin{pmatrix} \cdot & a & b & a \\ a & \cdot & a & b \\ b & a & \cdot & a \\ a & b & a & \cdot \end{pmatrix}.$$

Definition 2.7. The **GTR** (General Time Reversible) model is the most general stationary time-reversible evolutionary model possible, since it assumes different rates of substitution for each pair of nucleotides, in addition to assuming different frequencies of occurrence of nucleotides. It does not have any restriction on $\boldsymbol{\pi}$ other than $\sum_i \pi_i = 1$ and the rate matrix has the form

$$Q = \begin{pmatrix} \cdot & a\pi_C & b\pi_G & c\pi_T \\ a\pi_A & \cdot & d\pi_G & e\pi_T \\ b\pi_A & d\pi_C & \cdot & f\pi_T \\ c\pi_A & e\pi_C & f\pi_G & \cdot \end{pmatrix}.$$

Definition 2.8. The **Discrete Gamma** model [X. et al., 1995] considers several rate categories approximating the gamma distribution with all rates having equal probability and each site having p_i probability of mutating at rate category i.

The following proposition states the eigenvector decomposition of a reversible rate matrix (cfr. Levin and Peres [2017]).

Proposition 2.9. *For an irreducible rate matrix $Q = (q_{ij})$ over an alphabet $\mathbb{A}$ of $K+1$ states, the following holds:*

(a) Matrix Q has eigenvalue $\lambda_0 = 0$ with algebraic multiplicity 1. A right eigenvector of λ_0 is $\mathbf{1}$, whose left eigenvector is the unique equilibrium distribution $\boldsymbol{\pi}^T$, defined by equation $\boldsymbol{\pi}^T Q = \boldsymbol{\pi}^T$. The rest of eigenvalues of Q are complex number with strictly negative real part. Finally, the exponential matrix e^{Qt} satisfies

$$e^{Qt} \to \mathbf{1}\boldsymbol{\pi}^T \text{ as } t \to \infty.$$

(b) If matrix Q is reversible, then it has real eigenvalues $0 > \lambda_1 \geq ... \geq \lambda_K$. Moreover, matrix Q has an orthogonal basis of right eigenvectors $\boldsymbol{v_k}$ and left eigenvectors $\boldsymbol{h_k^T}$ such that $\boldsymbol{h_k} = \boldsymbol{\pi} \circ \boldsymbol{v_k} = Diag(\boldsymbol{\pi})\boldsymbol{v_k}$ for $k \in [0, K]$, where $\boldsymbol{v_0} = \mathbf{1}$, $\boldsymbol{h_0} = \boldsymbol{\pi}$ and $Q = \mathbf{1}\boldsymbol{\pi}^T + \boldsymbol{v_1}\boldsymbol{h_1^T}\lambda_1 + ... + \boldsymbol{v_K}\boldsymbol{h_K^T}\lambda_K$. Finally, the exponential matrix e^{Qt} can be computed as

$$e^{Qt} = \mathbf{1}\boldsymbol{\pi}^T + \boldsymbol{v_1}\boldsymbol{h_1^T}e^{\lambda_1 t} + ... + \boldsymbol{v_K}\boldsymbol{h_K^T}e^{\lambda_K t}.$$

2.3 Likelihood in a Phylogenetic Tree

In this section we describe the likelihood computation in a phylogenetic tree, introduced by Felsenstein [2004]. The likelihood of a tree is the probability of obtaining the observed sequence alignment given a tree topology, branch lengths and substitution model.

Given the observed data D, a set of aligned sequences of length m, the likelihood of an evolutionary process E is $\Pr(D|E)$. Because the evolution in different sites is independent, we can decompose the likelihood into a product, one term for each site, as

$$L(E) = \Pr(D|E) = \prod_{i=1}^{m} \Pr(D^{(i)}|E), \tag{1}$$

where $D^{(i)}$ is the data at the i'th site.

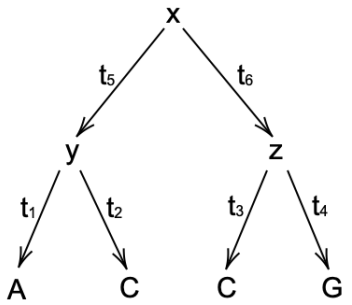


Figure 2: Tree with branch lengths and data at a single site.

Suppose that we have a tree with the data at a single site as shown in Figure 2. Then, the likelihood of the process given this site is the sum of the probabilities of each scenario of events over all the possible nucleotides $\mathbb{A} = \{A, C, G, T\}$ at the interior nodes of the tree, namely

$$\begin{aligned}\Pr(D^{(i)}|E) = \sum_{x\in\mathbb{A}}\sum_{y\in\mathbb{A}}\sum_{z\in\mathbb{A}} \Pr(x)\Pr(y|x,t_5)\Pr(z|x,t_6) \\ \Pr(A|y,t_1)\Pr(C|y,t_2)\Pr(C|z,t_3)\Pr(G|z,t_4).\end{aligned} \tag{2}$$

The assumption that evolution is independent in different lineages allows us to decompose $\Pr(D^{(i)}|E)$ as

$$\begin{aligned}\Pr(D^{(i)}|E) = \sum_{x\in\mathbb{A}} \Pr(x) \left(\sum_{y\in\mathbb{A}} \Pr(y|x,t_5)\Pr(A|y,t_1)\Pr(C|y,t_2)\right) \\ \left(\sum_{z\in\mathbb{A}} \Pr(z|x,t_6)\Pr(C|z,t_3)\Pr(G|z,t_4)\right),\end{aligned} \tag{3}$$

with the quantities inside the parentheses representing the likelihood of each of the subtrees.

Since the observed data at a single site i, $D^{(i)}$, forms a pattern, we can rewrite Equation 1 as

$$L(E) = \prod_{\partial} \Pr(\partial|E)^{n_\partial}, \tag{4}$$

where n_∂ is the number of times a pattern ∂ is observed in the data D. Hence, the log-likelihood of process E is

$$\mathcal{L}(E) = \sum_{\partial} n_\partial \log(\Pr(\partial|E)). \tag{5}$$

To estimate the true process that generated the data, we use a maximum likelihood estimate [Springer Verlag GmbH, European Mathematical Society, 2022a].

The Pulley principle [Felsenstein, 1981] states that, if the evolutionary process is reversible and stationary, then the root of the tree cannot be identified. This is due to the fact that the replacement of the root gives the same probability of observing pattern ∂ assuming process E. Consequently, the diagrams of Figure 3 are equivalent.

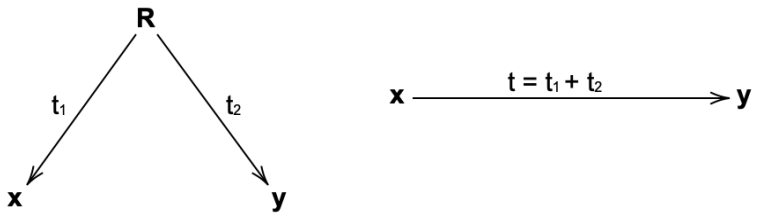


Figure 3: Two equivalent stationary and reversible processes as stated by the Pulley principle. On the left, the child configurations x and y are obtained by mutating the root sequence R as determined by the rate matrix Q. On the right, the root sequence x is sampled using the stationary distribution $\boldsymbol{\pi}$ and y is generated by mutating it according to Q.

If we consider an evolutionary process E on a tree rooted at R, the likelihood vector of R given pattern ∂ is the vector $\boldsymbol{\rho_\partial}$ of probabilities of observing ∂ at one site given each possible nucleotide at that site of the root, that is,

$$\boldsymbol{\rho_\partial} := (\rho_\partial^i) := (\Pr(\partial | i \text{ at node } R)), \tag{6}$$

where $i \in \mathbb{A}$ and E is omitted for simplicity. Since $\sum_\partial \Pr(\partial |\ i \text{ at node } R) = 1$ for all $i \in \mathbb{A}$, meaning that the sum of probabilities of observing all possible patterns given a nucleotide at the root is 1, it follows that $\sum_\partial \boldsymbol{\rho_\partial} = \mathbf{1}$.

Moreover, using the law of total probability [Springer Verlag GmbH, European Mathematical Society, 2022b], the likelihood of a process E given pattern ∂ is

$$L(E) = \Pr(\partial | E) = \sum_{i \in \mathbb{A}} \Pr(i \text{ at node } R) \Pr(\partial | i \text{ at node } R) = \boldsymbol{\pi} \cdot \boldsymbol{\rho_\partial}, \tag{7}$$

where $\cdot$ is the Euclidean product.

Now consider two nodes A and B adjacent to a common root R, with two subtrees rooted at nodes A and B, called clades, as showed in Figure 4. The evolutionary distance from R to nodes A and B is t_1 and t_2, respectively, with $t_1 + t_2 = t$. If the tree has a total of m leaves, now we have a clade A with k leaves and a clade B with $m - k$ leaves. A pattern ∂ induces sub-patterns ∂A and ∂B. Conversely, the subpatterns ∂A and ∂B determine pattern ∂.

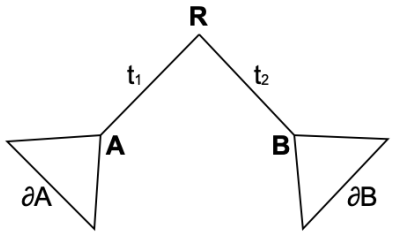


Figure 4: Diagram of the two clades rooted at A and B, separated by branch AB and with evolutionary distance t_1 and t_2, respectively, to the common root R. Pattern ∂ induces subpatterns ∂A and ∂B.

If we can observe only pattern ∂A and the process consists only on clade A, the likelihood vector at node A given ∂A is

$$\boldsymbol{\alpha_{\partial A}} := (\alpha^i_{\partial A}) := (\Pr(\partial A | i \text{ at node } A)). \tag{8}$$

Analogously, the likelihood vector at node B given ∂B is

$$\boldsymbol{\beta_{\partial B}} := (\beta^i_{\partial B}) := (\Pr(\partial B | i \text{ at node } B)). \tag{9}$$

Since the probability of state $i \in \mathbb{A}$ mutating to $j \in \mathbb{A}$ in time t is $p_{ij}(t)$, where $(p_{ij}(t)) = e^{Qt}$, the likelihood vector at node R given only clade A is

$$\Pr(\partial A | i \text{ at node } R) = e^{Qt_1}\boldsymbol{\alpha_{\partial A}}. \tag{10}$$

Analogously, the likelihood vector at node R given only clade B is

$$\Pr(\partial B | i \text{ at node } R) = e^{Qt_2}\boldsymbol{\beta_{\partial B}}. \tag{11}$$

Since subpatterns ∂A and ∂B are a partition of ∂, we have

$$\boldsymbol{\rho_\partial} = (e^{Qt_1}\boldsymbol{\alpha_{\partial A}}) \circ (e^{Qt_2}\boldsymbol{\beta_{\partial B}}), \tag{12}$$

where $\circ$ is the element-wise product, also known as Hadamard product.

Because we are considering a reversible and stationary process, we can place the root on node A without altering the likelihood. Therefore we set $t_1 = 0$ and $t_2 = t$, giving

$$\boldsymbol{\rho_\partial} = \boldsymbol{\alpha_{\partial A}} \circ (e^{Qt}\boldsymbol{\beta_{\partial B}}). \tag{13}$$

All in all, given pattern ∂, the likelihood of branch AB having length t is

$$\Pr(\partial|t) = \boldsymbol{\pi} \cdot \boldsymbol{\rho_\partial} = \boldsymbol{\pi} \cdot \left(\boldsymbol{\alpha_{\partial A}} \circ (e^{Qt}\boldsymbol{\beta_{\partial B}})\right) = \boldsymbol{\alpha}_{\boldsymbol{\partial A}}^T \mathrm{Diag}(\boldsymbol{\pi}) e^{Qt} \boldsymbol{\beta_{\partial B}}. \tag{14}$$

All multiples of the likelihood vector can be used to have a maximum likelihood estimator of the ancestor identity at the root. The normalised likelihood vector at the root R was defined by Manuel [2022] in order to have a unique representation of all likelihood vectors as

$$\tilde{\boldsymbol{\rho}}_{\boldsymbol{\partial}} := \frac{\boldsymbol{\rho_\partial}}{\boldsymbol{\pi} \cdot \boldsymbol{\rho_\partial}} = \frac{\boldsymbol{\rho_\partial}}{\Pr(\partial)}. \tag{15}$$

If $\boldsymbol{\rho_\partial}$ has nearly uniform entries, then it is difficult to estimate the ancestor sequence at the root R, because the probability of observing the pattern ∂ would be approximately equal given each possible nucleotide at the root. Moreover, vector $\boldsymbol{\rho_\partial}$ is uniform when $\tilde{\boldsymbol{\rho}}_{\boldsymbol{\partial}} = \mathbf{1}$, where $\mathbf{1}$ is the column vector full of 1's.

2.3.1 Decomposition of the likelihood of a branch

Here we introduce the spectral decomposition of the likelihood of a branch as described by Manuel [2022].

Assuming that Q is reversible and using Proposition 2.9.b, we have $\mathrm{Diag}(\boldsymbol{\pi})e^{Qt} = \boldsymbol{\pi}\boldsymbol{\pi}^T + \sum_{k\in[K]} \boldsymbol{h_k}\boldsymbol{h_k}^T e^{\lambda_k t}$. Thus we can then rewrite Equation 14 as

$$\Pr(\partial|t) = \boldsymbol{\alpha}_{\boldsymbol{\partial A}}^T \mathrm{Diag}(\boldsymbol{\pi}) e^{Qt} \boldsymbol{\beta_{\partial B}} = \Pr(\partial A)\Pr(\partial B) + \sum_{k\in[K]} (\boldsymbol{\alpha_{\partial A}} \cdot \boldsymbol{h_k})(\boldsymbol{\beta_{\partial B}} \cdot \boldsymbol{h_k}) e^{\lambda_k t}. \tag{16}$$

Equation 16 shows that $\Pr(\partial|t) = \Pr(\partial A)\Pr(\partial B)$ when subpatterns ∂A and ∂B are independent, since the second term of Equation 16 vanishes for large t because λ_k is always negative. Therefore, the dependence factor $D(\partial|t)$ can be defined to measure the dependence between events ∂A and ∂B as

$$D(\partial|t) := \frac{\Pr(\partial|t)}{\Pr(\partial A)\Pr(\partial B)} = \tilde{\boldsymbol{\alpha}}_{\boldsymbol{\partial A}}^T \mathrm{Diag}(\boldsymbol{\pi}) e^{Qt} \tilde{\boldsymbol{\beta}}_{\boldsymbol{\partial B}}, \tag{17}$$

where we used Equation 15. The dependence factor will be equal to 1 if the two subpatterns are independent. Using Equation 16, the dependence factor can be decomposed as

$$D(\partial|t) = 1 + \sum_{k\in[K]} (\tilde{\boldsymbol{\alpha}}_{\boldsymbol{\partial A}} \cdot \boldsymbol{h_k})(\tilde{\boldsymbol{\beta}}_{\boldsymbol{\partial B}} \cdot \boldsymbol{h_k}) e^{\lambda_k t}. \tag{18}$$

Note that $D(\partial|t) \to 1$ as $t \to \infty$, indicating again that subpatterns ∂A and ∂B become independent as the branch length t grows.

3 New Measures of Phylogenetic Information

3.1 Coherence

In this section we describe the concept of coherence, introduced by Manuel and von Haeseler [2022].

Given two vectors $\boldsymbol{v} = (v_i)$ and $\boldsymbol{w} = (w_i)$, their π-inner product is

$$\langle \boldsymbol{v}, \boldsymbol{w} \rangle_\pi := \boldsymbol{\pi} \cdot (\boldsymbol{v} \circ \boldsymbol{w}) = \boldsymbol{v}^T \mathrm{Diag}(\boldsymbol{\pi}) \boldsymbol{w} = \sum_i \pi_i v_i w_i. \tag{19}$$

Consider two adjacent nodes A and B determining branch AB, which induces a partition of ∂ into ∂A and ∂B. Using the π-inner product, the **coherence** of branch AB given ∂ is defined as

$$C^\partial(A; B) := \langle \tilde{\boldsymbol{\rho}}_{\boldsymbol{\partial A}} - \mathbf{1}, \tilde{\boldsymbol{\rho}}_{\boldsymbol{\partial B}} - \mathbf{1} \rangle_\pi = \langle \tilde{\boldsymbol{\rho}}_{\boldsymbol{\partial A}}, \tilde{\boldsymbol{\rho}}_{\boldsymbol{\partial B}} \rangle_\pi - 1, \tag{20}$$

using that $\langle \tilde{\boldsymbol{\rho}}_{\boldsymbol{\partial A}}, \mathbf{1} \rangle_\pi = \langle \tilde{\boldsymbol{\rho}}_{\boldsymbol{\partial B}}, \mathbf{1} \rangle_\pi = 1$.

The **population coherence** of branch AB is defined as

$$C(A; B) := \mathbb{E}[C^\partial(A; B)] = \sum_\partial \Pr(\partial) C^\partial(A; B). \tag{21}$$

In words, the population coherence is the sum of the coherence of all patterns multiplied by their respective probabilities of being observed. The population coherence quantifies the dependence between clades A and B, because it tends to zero as the true length of branch AB grows.

Given an alignment of n sites where pattern ∂ is observed n_∂ times, the **sample coherence** of branch AB is the coherence of a representative part of the population and is defined as

$$\hat{C}(A; B) := \sum_\partial \frac{n_\partial}{n} C^\partial(A; B). \tag{22}$$

3.1.1 Decomposition of the coherence

The coherence between nodes A and $e^{Qt}B$ is

$$C^\partial(A; e^{Qt}B) = \langle \tilde{\boldsymbol{\alpha}}_{\boldsymbol{\partial A}}, e^{Qt} \tilde{\boldsymbol{\beta}}_{\boldsymbol{\partial B}} \rangle_\pi - 1 = \tilde{\boldsymbol{\alpha}}^T_{\boldsymbol{\partial A}} \mathrm{Diag}(\boldsymbol{\pi}) e^{Qt} \tilde{\boldsymbol{\beta}}_{\boldsymbol{\partial B}} - 1 = D(\partial | t) - 1. \tag{23}$$

Thus, the coherence of a branch AB goes to zero as subpatterns ∂A and ∂B are close to independent.

Using Equation 18 and the fact that $C^{\partial}(A; e^{Qt}B) = C^{\partial}(A; B)$ when $t = 0$, one can write

$$C^{\partial}(A; B) = \sum_{k \in [K]} (\tilde{\boldsymbol{\alpha}}_{\boldsymbol{\partial A}} \cdot \boldsymbol{h_k})(\tilde{\boldsymbol{\beta}}_{\boldsymbol{\partial B}} \cdot \boldsymbol{h_k}). \tag{24}$$

Motivated by this decomposition, the k-projection of the coherence of branch AB given pattern ∂ is defined as

$$C^{\partial}_k(A; B) := (\tilde{\boldsymbol{\alpha}}_{\boldsymbol{\partial A}} \cdot \boldsymbol{h_k})(\tilde{\boldsymbol{\beta}}_{\boldsymbol{\partial B}} \cdot \boldsymbol{h_k}). \tag{25}$$

Hence, Equation 24 can be rewritten as

$$C^{\partial}(A; B) = \sum_{k \in [K]} C^{\partial}_k(A; B). \tag{26}$$

The k-projection of the population coherence of branch AB is defined as

$$C_k(A; B) := \mathbb{E}[C^{\partial}_k(A; B)]. \tag{27}$$

and the k-projection of the sample coherence of branch AB as

$$\hat{C}_k(A; B) := \sum_{\partial} \frac{n_{\partial}}{n} C^{\partial}_k(A; B). \tag{28}$$

3.2 Memory

In this section we explain the concept of memory, introduced in Manuel and von Haeseler [2022].

In Equation 15 the memory vector is defined as $\tilde{\boldsymbol{\rho}}_{\boldsymbol{\partial}} - \mathbf{1}$, meaning that it would be zero when $\boldsymbol{\rho}_{\boldsymbol{\partial}}$ is uniform and increase as the ancestor sequence at the root R becomes more identifiable. Hence, the module of the memory vector can be used to describe the expected "amount of identification" of the ancestral sequence.

Given the π-inner product defined in 19, the $L_2(\pi)$-norm of a vector $\boldsymbol{v}$ is $\|\boldsymbol{v}\|_{\pi} = \sqrt{\langle \boldsymbol{v}, \boldsymbol{v} \rangle_{\pi}} = \sqrt{\sum_i \pi_i v_i^2}$. From this, the **memory** of a clade R given pattern ∂ is defined as

$$M^{\partial}(R) := \|\tilde{\boldsymbol{\rho}}_{\boldsymbol{\partial}} - \mathbf{1}\|^2_{\pi}, \tag{29}$$

while the **population memory** of a clade can be defined as

$$M(R) := \mathbb{E}[M^{\partial}(R)] = \sum_{\partial} \Pr(\partial) M^{\partial}(R). \tag{30}$$

The population memory is an average for all patterns of $\|\tilde{\boldsymbol{\rho}}_{\boldsymbol{\partial}} - \mathbf{1}\|_{\pi}^2$. It is small if the likelihood of the ancestral states at the root is uniform. Therefore, it quantifies the identification of the root, that is, how confidently we expect to reconstruct the root ancestral state.

If there is an alignment where pattern ∂ is observed n_{∂} times, to estimate $M(R)$ we use the **sample memory** of clade R, which is defined as

$$\hat{M}(R) := \sum_{\partial} \frac{n_{\partial}}{n} M^{\partial}(R). \tag{31}$$

3.2.1 Decomposition of the memory

Algebraically, for a clade R, $M^{\partial}(R) = C^{\partial}(R; R)$. Then, using Equation 24 one can write

$$M^{\partial}(R) = \sum_{k \in [K]} (\tilde{\boldsymbol{\rho}}_{\boldsymbol{\partial}} \cdot \boldsymbol{h}_{\boldsymbol{k}})^2. \tag{32}$$

The kl-projection of the memory of a clade R given ∂ is defined as

$$M^{\partial}_{kl}(R) := (\tilde{\boldsymbol{\rho}}_{\boldsymbol{\partial}} \cdot \boldsymbol{h}_{\boldsymbol{k}})(\tilde{\boldsymbol{\rho}}_{\boldsymbol{\partial}} \cdot \boldsymbol{h}_{\boldsymbol{l}}), \tag{33}$$

and thus we can rewrite Equation 32 as

$$M^{\partial}(R) = \sum_{k \in [K]} M^{\partial}_{kk}(R) \tag{34}$$

The kl-projection of the population memory of a clade R is defined as

$$M^{\partial}_{kl}(R) := \mathbb{E}[M^{\partial}_{kl}(R)], \tag{35}$$

and the kl-projection of the sample memory of a clade R as

$$\hat{M}_{kl}(R) := \sum_{\partial} \frac{n_{\partial}}{n} M^{\partial}_{kl}(R). \tag{36}$$

3.3 Branch Saturation

Saturation is defined by Manuel and von Haeseler [2022] as the lack of significance to reject the null hypothesis that the alignment was generated from an infinite evolutionary

process. Following this definition, we say a branch AB is saturated if we cannot reject the null hypothesis that the true length t^* of the branch is infinite. Recall that $t^* \to \infty$ implies that subpatterns ∂A and ∂B are independent. Since we assume a reversible process, the root of any tree is unidentifiable and, therefore, if branch AB is saturated, then the reconstructed tree is composed by unrooted clade A (removing parent node A) independent from unrooted clade B (removing parent node B).

A saturated branch AB can be explained in three ways:

- Too many mutations have happened due to t^* being too large. An alignment with more sites has a higher probability of rejecting saturation.
- Some sites are wrongly aligned.
- The assumed evolutionary process is incorrect. Either the tree topology, the branch lengths or the rate matrix are erroneous.

3.3.1 Dominant sample coherence

Manuel and von Haeseler [2022] define the dominant sample coherence as follows.

Proposition 3.1. *Given an alignment where pattern ∂ is observed n_∂ times, assume that the alignment is the realization of a stationary and reversible process on a tree with rate matrix Q. If matrix Q has eigenvalues $0 \geq \lambda_1 \geq ... \geq \lambda_K$ such that λ_1 has multiplicity $D \leq K$, then the dominant sample coherence can be defined as*

$$\hat{\delta} := \sum_{k\in[D]} \hat{C}_k(A;B) = \sum_{k\in[D]} \sum_{\partial} \frac{n_\partial}{n} (\boldsymbol{h_k} \cdot \tilde{\boldsymbol{\alpha}}_{\boldsymbol{\partial A}})(\boldsymbol{h_k} \cdot \tilde{\boldsymbol{\beta}}_{\boldsymbol{\partial B}}).$$

Assuming $t \to \infty$, the dominant sample coherence $\hat{\delta}$ satisfies that

$$\mathbb{E}[\hat{\delta}|t^* \to \infty] = \sum_{k\in[D]} \mathbb{E}[C_k^\partial(A;B)|t^* \to \infty] = 0. \tag{37}$$

Since we assume that the sites of an alignment are generated independently, the integers n_∂ are multinomially distributed with probabilities $\Pr(\partial)$. Therefore, for large n, each observed quantity n_∂ can be approximated as the outcome a normal distribution, as well as the sample dominant coherence $\hat{\delta}$, which is a linear combination of the integers n_∂. Thus, the distribution of statistic $\hat{\delta}$ can be approximated as

$$\hat{\delta} \sim N(0, \sigma^2), \tag{38}$$

where the variance σ^2 is defined as

$$\sigma^2 := Var[\hat{\delta}|t^* \to \infty] = \frac{1}{n} \sum_{k,l\in[D]} M_{kl}(A) M_{kl}(B), \tag{39}$$

and $N(\mu, \sigma^2)$ is the normal distribution [Springer Verlag GmbH, European Mathematical Society, 2022c].

3.3.2 The asymptotic test for branch saturation

In order to detect if a branch is saturated or not, Manuel and von Haeseler [2022] propose, given a level of significance α, the following asymptotic test:

$$\text{"Reject } t^* \to \infty \quad \text{if} \quad \hat{\delta} > c_s\text{"}, \tag{40}$$

where the saturation coherence $c_s \in [0, 1]$ is chosen so that

$$\Pr(\hat{\delta} > c_s | t^* \to \infty) = \alpha. \tag{41}$$

When $\hat{\delta} > c_s$ we cannot reject the null hypothesis that $t^* \to \infty$ and, hence, we say that branch AB is saturated with significance α. Otherwise, we say that branch AB is informative with significance α.

From Equations 38 and 39, we can approximate c_s for large n as

$$c_s \approx z_\alpha \sqrt{\frac{1}{n} \sum_{k,l \in [D]} M_{kl}(A) M_{kl}(B)}, \tag{42}$$

where z_α is defined such that $\Pr(Z > z_\alpha) = \alpha$ for $Z \sim N(0, 1)$.

The asymptotic test is simplified if $D = 1$. Then $\hat{\delta} = \hat{C}_1(A; B)$ and

$$c_s \approx z_\alpha \sqrt{\frac{M_{11}(A) M_{11}(B)}{n}}, \tag{43}$$

The asymptotic test is also simplified if the branch AB is an external branch, meaning that either A or B is a leaf. Assuming, for example, that A is a leaf, then

$$c_s \approx z_\alpha \sqrt{\frac{1}{n} \sum_{k \in [D]} M_{kk}(B)}. \tag{44}$$

In addition, if B is also a leaf, then the saturation coherence is

$$c_s \approx z_\alpha \sqrt{\frac{D}{n}}. \tag{45}$$

If a clade A has a small number of leaves, then it is possible to compute $M_{kl}(A)$ numerically. In more complicated instances, we build confidence intervals of radius ϵ around each sample memory $\hat{M}_{kl}(A)$ and $\hat{M}_{kl}(B)$ using the fact that $\mathrm{Var}[\hat{M}^{\partial}_{kl}(R)] \leq \sqrt{\frac{U\min(K,U/4)}{n}}$. Then we use ϵ to construct a confidence interval around $\sum_{k,l} \hat{M}_{kl}(A)\hat{M}_{kl}(B)$. This slightly increases c_s and improves the stability of the test.

4 Implementation

Figure 5 shows a summary of our implementation of the saturation test.

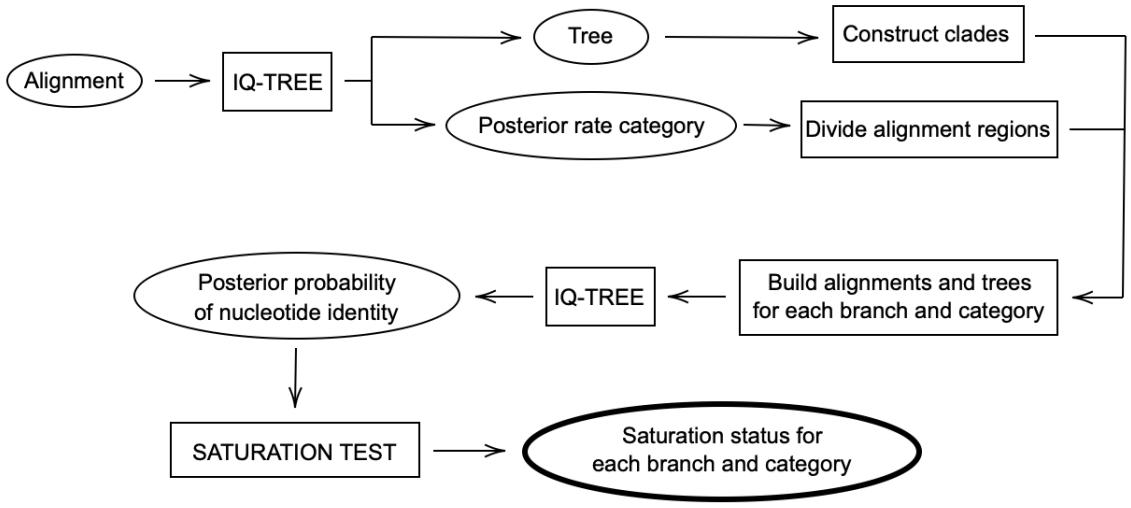


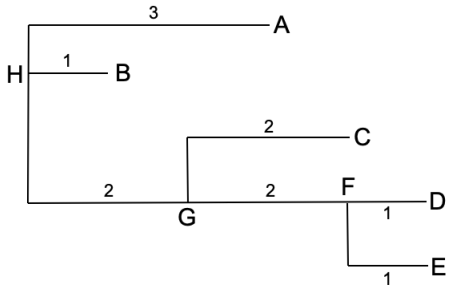
Figure 5: Diagram representing the workflow of the implementation.

Our input is a multiple sequence alignment. Then, IQ-TREE [B.Q. et al., 2020], using the maximum likelihood criterion, estimates the (hopefully) optimal substitution model and evolutionary tree given the alignment.

The output that we need from IQ-TREE is the tree topology and the branch lengths. Moreover, assuming the discrete Gamma model (Definition 2.8), then we need the probability of each site mutating at each of the different rates, as also the rates themselves.

IQ-TREE outputs its reconstructed tree in Newick format [Felsenstein, 2004, Olsen G., 1990], a way of representing graph trees using parentheses and commas. In Newick format, the distance from a node to its parent is indicated by a number following a colon written at the right side of the node's name. Nodes that have a common parent node are separated by commas. Those nodes written inside the same parenthesis are part of the same subtree and the node which is written at the right side of a closing parenthesis is the parent of that subtree. Typically, a tree's representation is rooted on an internal node. When an

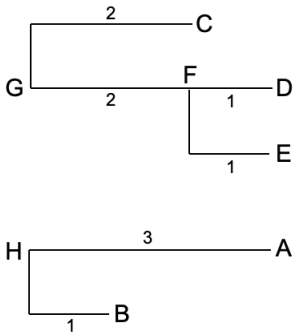
unrooted tree is represented in Newick notation, an arbitrary node is chosen as root. A simple tree and its corresponding Newick format is shown in Figure 6.



(A:3,B:1,(C:2,(D:1,E:1)F:2)G:2)H;

Figure 6: Tree (left) and its corresponding Newick format (right). In this case the tree is rooted at node H. The vertical distances are meaningless.

Once we have the output from IQ-TREE, we need to read its Newick format. Then we split the two clades connected by each branch and store the clades. For instance, the split induced by branch HG in Figure 6 is shown in Figure 7.



(C:2,(D:1,E:1)F:2)G;

(A:3,B:1)H;

Figure 7: Trees (left) and their corresponding Newick formats (right) of the two subtrees connected by branch HG.

Assuming the discrete Gamma model, we separate the alignment into regions, each corresponding to its most likely rate category. To do so, we use the IQ-TREE output containing the probabilities of each site mutating at each of the different rates. For a given site, if all rate categories are equally likely, we consider that site as uninformative and remove it. Then we store copies of all the clades with their branch lengths multiplied by each of the rates. For each clade and rate category, we store their corresponding subalignment region.

We input each clade and its corresponding subalignment region into IQ-TREE, while fixing the same evolutionary process as the first IQ-TREE reconstruction. This gives, for each site, the posterior probability $\boldsymbol{r_\partial} = (r_A, r_C, r_G, r_T)$ of observing each nucleotide at the parent node of the clade. Since we assume a stationary process, the posterior $\boldsymbol{r_\partial}$ is closely related to the normalized likelihood vector $\tilde{\boldsymbol{\rho}}_{\boldsymbol{\partial}}$ through the relationship

$$\tilde{\boldsymbol{\rho}}_{\boldsymbol{\partial}} \circ \boldsymbol{\pi} = \boldsymbol{r_\partial}. \tag{46}$$

We diagonalise the rate matrix to obtain the largest non-zero eigenvalues and their corresponding eigenvectors, which will be used for the asymptotic test.

Finally, we compute the asymptotic test for branch saturation. To do so, we need to compute the dominant sample coherence $\hat{\delta}$ of Proposition 3.1 for each branch and rate category. Then, we calculate the saturation coherence c_s of Equation 42. A confidence interval of radius ϵ is added to c_s, as explained in Section 3.3.2.

The final output of the implementation are the values of the statistic $\hat{\delta}$ and the saturation coherence c_s for each branch and rate category. From this, we can determine the saturation status of each branch assuming each rate category: either the branch is saturated (when $\hat{\delta} < c_s$) or informative (when $\hat{\delta} > c_s$).

The Python code of the implementation can be acccessed in `https://drive.google.com/drive/folders/1UaqdUyGxb99Di-N34h9Bs4CYuYtziVkw?usp=sharing`.

5 Results

5.1 Example 1

In this example we use the DNA alignment of length 3265 of the ENV gene from five SIV sequences, extracted from LANL [2020] [Foley et al., 2020]. Due to their high rate of mutation, phylogenies of SIV tend to vary depending on the data used for the analysis [K. et al., 2009]. Hence, we aim to use the asymptotic test to detect if a region of the alignment is not supporting a particular branch of the tree.

We use IQ-TREE assuming a GTR model and a Gamma model with four mutation rates, to reconstruct the maximum likelihood tree of these SIV species. The reconstructed tree is shown in Figure 8.

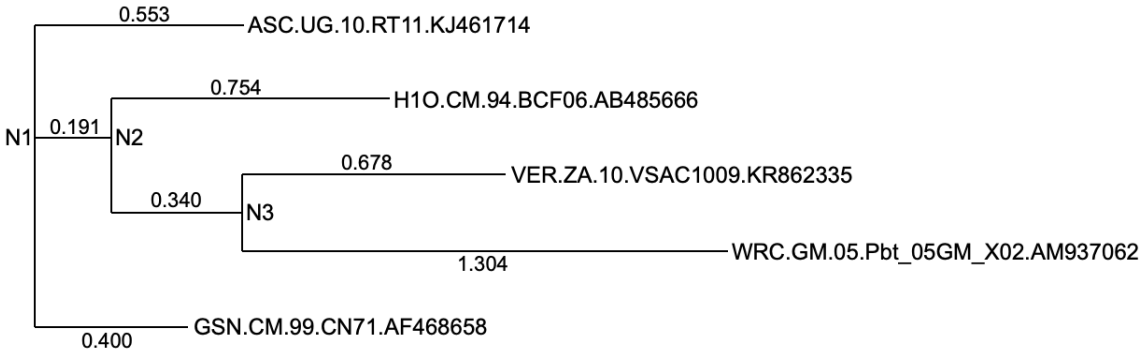


Figure 8: Reconstructed phylogenetic tree of the five SIV species with labelled nodes and branch lengths. There are three internal nodes, called N1, N2 and N3.

For simplification, let us call the different SIV sequences SIV1, SIV2, SIV3, SIV4 and SIV5, corresponding to the sequences shown from top to bottom in Figure 8.

According to the IQ-TREE reconstruction, the four substitution rates are $f_1 = 0.2087$, $f_2 = 0.5817$, $f_3 = 1.077$ and $f_4 = 2.273$. The alignment regions corresponding to each of the four rates have lengths $n_1 = 991$, $n_2 = 431$, $n_3 = 390$ and $n_4 = 1008$. As the original branch lengths are multiplied by each of the four rates, we have four trees with different branch lengths.

We choose significance level $\alpha = 0.05$, giving $z_\alpha \approx 1.96$, to execute the asymptotic test to detect branch saturation.

In the two regions corresponding to the f_1 and f_2, the test rejects the null hypothesis in all branches. This is reasonable, since their corresponding modified branches are considerably shorter than the original branches.

In the alignment region corresponding to the f_3, the null hypothesis that $t^* \to \infty$ cannot be rejected in the longest branch of the tree, connecting node N3 with the leaf corresponding to the sequence SIV5, which has length 1.404 after being multiplied by the corresponding mutation rate. We can say that the region with rate f_3 has not collaborated in the reconstruction of this branch. In this case, $\hat{\delta} = 0.04$ and $c_s = 0.098$. For this rate, the rest of branches are informative.

In the region with reconstructed rate f_4 the asymptotic test cannot reject the null hypothesis in any of the branches and concludes that all of them are saturated. In this case, the dominant sample coherence $\hat{\delta}$ takes negative values around -0.1 in the external branches and -0.07 in the internal ones, while c_s is around 0.03. Intuitively, we may say that the IQ-TREE reconstruction method has grouped under the largest rate all uninformative patterns. Consequently, this region, which has about one third of the sites, can be ignored without significantly affecting the phylogeny, or equivalently we could just set $f_4 = \infty$. All in all, the asymptotic test for saturation has recognized a region not

providing significant information for the reconstructed phylogeny.

5.2 Example 2

In this example we use a simulated binary and fully balanced tree (with left and right subtrees of any node having the same amount of leaves) with 64 sequences of size $n = 1000$. The species can be grouped in four subtrees of 16 taxa each, with all the branches in these clades having length 0.01. The clades are connected by branches of length 3, and those branches are connected by an internal branch of length 0.02. A simplified representation of this example is shown in Figure 9.

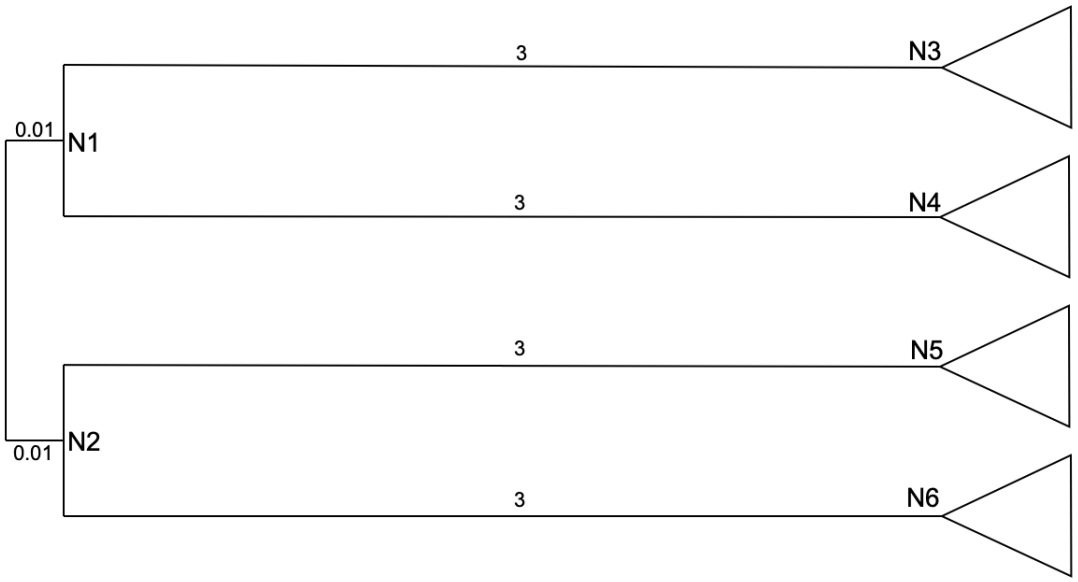


Figure 9: Simulated binary and fully balanced tree with 64 taxa grouped in four clades represented by four triangles. The branches' lengths are shown. The internal nodes are called N1, N2, N3, N4, N5 and N6.

We use IQ-TREE assuming a JC model, in addition to a Gamma model with four mutation rates. In this case, we fix the tree and branch lengths.

According to the IQ-TREE reconstruction, the four substitution rates are $f_1 = 0.1347$, $f_2 = 0.4733$, $f_3 = 0.9981$ and $f_4 = 2.394$. The alignment regions corresponding to each of the four rates have lengths $n_1 = 191$, $n_2 = 235$, $n_3 = 255$ and $n_4 = 319$. As the original branch lengths are multiplied by each of the four rates, we have four trees with different branch lengths.

To execute the asymptotic test to detect branch saturation, we choose significance level $\alpha = 0.05$.

The branches inside the four clades are always informative for all substitution rates, meaning that a significant amount of information is conserved from the leaves to nodes N3,

N4, N5 and N6.

In the regions corresponding to rate f_1 the test rejects the null hypothesis in all branches and, hence, we say that all of them are informative.

In the regions corresponding to rates f_2, f_3 and f_4 the four long branches are saturated, since the test cannot reject the null hypothesis in any of them. Consequently, when it comes to these three regions, the information carried from the leaves is lost on nodes N3, N4, N5 and N6. Hence, the four clades of 16 taxa would be independent.

Moreover, in the three regions corresponding to f_2, f_3 and f_4, the internal short branch connecting nodes N1 and N2 is saturated too. This gives the intuition that the saturation of a branch does not depend its length, but on the information at the two nodes connected by the branch. Thus, we can say that a short branch is not necessarily informative, specially if such branch is internal.

5.3 Example 3

In this example we use the same 64 simulated sequences of size n = 1000 as in Section 5.2, but without fixing the model, the tree topology nor the branch lengths, allowing IQ-TREE to determine the optimal model and tree by the maximum likelihood criterion. IQ-TREE selects JC as model, in addition to a discrete Gamma. The maximum likelihood tree is shown in Figure 10.

While the simulated tree in Figure 9 has four long branches of equal length connecting four clades of 16 taxa, the maximum likelihood tree in Figure 10 has, as well, four long branches connecting four clades of 16 taxa, but of different lengths. Furthermore, in Figure 9 node N1 connects nodes N3 and N4, and node N2 connects nodes N5 and N6. In the reconstructed tree in Figure 10, however, node N1 connects nodes N3 and N6, and node N2 connects nodes N5 and N4. The taxa inside each clade is the same in both the simulated tree (Figure 9) and the reconstructed tree (Figure 10).

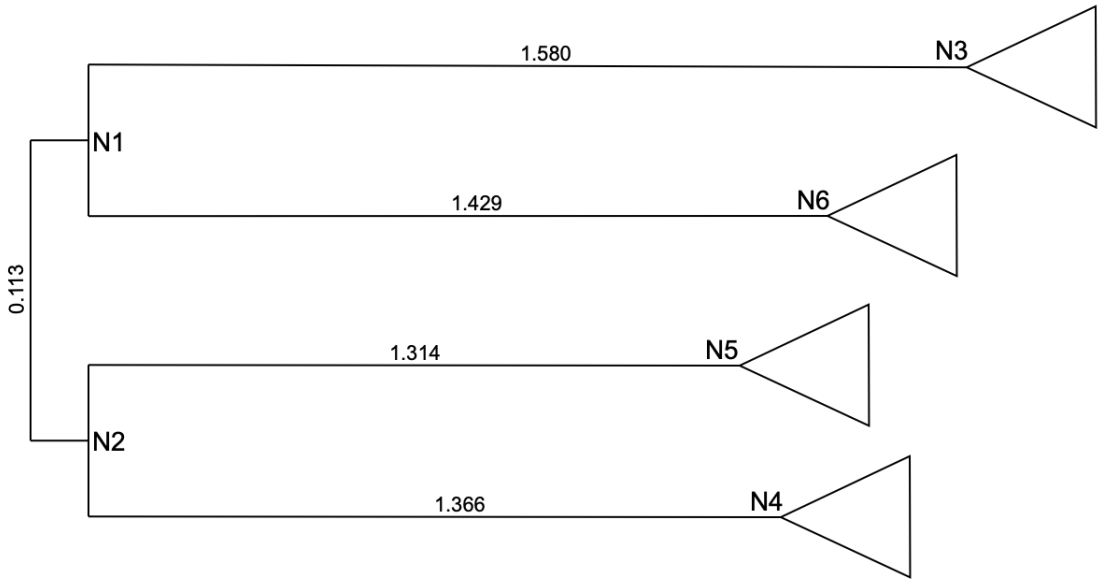


Figure 10: Tree with 64 taxa grouped in four clades represented by four triangles. The branches' lengths are shown. The internal nodes are called N1, N2, N3, N4, N5 and N6.

The four substitution rates are $f_1 = 0.1347$, $f_2 = 0.4733$, $f_3 = 0.9981$ and $f_4 = 2.394$. The alignment regions corresponding to each of the four rates have lengths $n_1 = 251$, $n_2 = 296$, $n_3 = 150$ and $n_4 = 303$. We have four trees with different branch lengths after multiplying the original branch lengths by each of the four rates.

To execute the asymptotic test to detect branch saturation, we choose significance level $\alpha = 0.05$.

The branches inside the four clades are always informative for all substitution rates.

In the regions corresponding to rates f_1 and f_2 the test rejects the null hypothesis in all branches and, hence, we say that all of them are informative.

In the regions corresponding to rates f_3 and f_4 the four long branches are saturated, since the test cannot reject the null hypothesis in any of them. Moreover, the internal short branch connecting nodes N1 and N2 is saturated too.

6 Conclusions

Our most important achievement has been the systematic implementation of the asymptotic test for branch saturation. Furthermore, we described and understood phylogenetic reconstruction using maximum likelihood, in addition to the new measures of phylogenetic introduced by Manuel and von Haeseler [2022].

From the test results on the three examples in Section 5 we can highlight the following conclusions:

- The sequence region corresponding to the fastest mutation rate tends to group all all uninformative patterns. Consequently, this region can be ignored without significantly affecting the phylogeny, which is equivalent to setting its corresponding rate to infinity. Thus the asymptotic test for saturation can recognize regions not providing significant information for the phylogenetic reconstruction.

- Long branches can be saturated even for small rates, as we have seen in Section 5.1.

- A short branch is not necessarily informative, specially for internal branches. The output of the test does not depend directly on the length of the branch, but on the information at the two nodes adjacent to such branch. We have seen this in Sections 5.2 and 5.3.

References

Minh B.Q., Schmidt H.A., Chernomor O., Schrempf D., Woodhams M.D, von Haeseler A., and Lanfear R. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution*, 37(5):1530–1534, 2020. https://doi.org/10.1093/molbev/msaa015.

J. Felsenstein. Evolutionary Trees from DNA Sequences: A Maximum Likelihood Approach. *Journal of Molecular Evolution*, 17:368–376, 1981.

J. Felsenstein. *Inferring Phylogenies*. Sinauer Associates, 2004.

B. Foley, T. Leitner, C. Apetrei, B. Hahn, I. Mizrachi, J. Mullins, A. Rambaut, S. Wolinsky, and B. Korber. HIV Sequence Compendium 2018. Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, NM, LA-UR 18-25673, 2020.

Strimmer K., Salemi M., and von Haeseler A. Genetic Distances and Nucleotide Substitution Models. In *The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing*, pages 111 – 141. Cambridge University Press, 2009. https://doi.org/10.1017/CBO9780511819049.006.

LANL. Los Alamos National Laboratory. http://www.hiv.lanl.gov/, 2020. Accessed: 2020-12-07.

D.A. Levin and Y. Peres. *Markov Chains and Mixing Times*, volume 107. American Mathematical Soc., 2017.

C. Manuel. An Upper Bound of the Information Flow From Children to Parent Node on Trees. 2022. https://doi.org/10.48550/arXiv.2204.10618.

C. Manuel and A. von Haeseler. New Measures of Phylogenetic Information Allow to Test for Saturation. *Unpublished*, 2022.

Olsen G. "Newick's 8:45" Tree Format Standard, 1990. http://evolution.genetics.washington.edu/phylip/newick_doc.html.

Springer Verlag GmbH, European Mathematical Society. Encyclopedia of Mathematics, 2022a. http://encyclopediaofmath.org/index.php?title=Maximum-likelihood_method&oldid=47805. Accessed on 2022-06-29.

Springer Verlag GmbH, European Mathematical Society. Encyclopedia of Mathematics, 2022b. http://encyclopediaofmath.org/index.php?title=Complete_probability_formula&oldid=46421. Accessed on 2022-08-29.

Springer Verlag GmbH, European Mathematical Society. Encyclopedia of Mathematics, 2022c. http://encyclopediaofmath.org/index.php?title=Normal_distribution&oldid=48011. Accessed on 2022-09-05.

Gu X., Fu Y.X., and Li W.H. Maximum Likelihood Estimation of the Heterogeneity of Substitution Rate among Nucleotide Sites. *Molecular Biology and Evolution*, 12(4): 546–557, 1995. https://doi.org/10.1093/oxfordjournals.molbev.a040235.