POWER OF THE LIKELIHOOD RATIO TEST FOR MODELS OF DNA BASE SUBSTITUTION

G.B. Cybis[a], S.R.C. Lopes[a1] and H.P. Pinheiro[b]

[a]Instituto de Matemática - UFRGS
Porto Alegre - RS - Brazil

[b]Instituto de Matemática e Estatística - UNICAMP
Campinas - SP - Brazil

January 25, 2011

Abstract

The goal of this work is to study the properties of the likelihood ratio tests comparing base substitution models. These are the most widely used hypothesis tests. With mild regularity conditions, we show that the asymptotic distribution of the likelihood ratio statistic test, under the alternative hypothesis, is a non-central Chi-square $\chi^2_k(D)$ distribution. The asymptotic normal distribution of the likelihood ratio test is proved when the sequence length S goes to infinity. We also propose a consistent estimator for the non-centrality parameter D. Through asymptotic theory and based on this consistent estimator for D, we propose a low computational cost estimator for the power of the likelihood ratio test. The methodology is applied to 17 different gene sequences of the ECP-EDN family in primates.

Keywords. Phylogenetic Inference; Monte Carlo Simulation; Evolutionary Model; Maximum Likelihood Function; Hypothesis Test; Power of the Test; Asymptotic Distribution.

1 Introduction

In the last few years the availability of sequenced DNA has increased dramatically. These gene sequences not only contain precious information about regular biological function and disease determination, but also bear witness to the evolutionary history of the group of organisms to which they belong. Evolutionary studies using phylogenetics (study of the evolutionary relationship between organisms) have been employed for a variety of purposes, including better definition of systematic classification, uncovering gene histories of duplication events (see Bielawski and Yang, 2003), dating the most recent common ancestor of clades (see Ho and Phillips, 2009) and inferring the presence of selective pressure on certain portions of the DNA (see Suzuki and Gojobori, 1999).

Genes and genomes are made with DNA, a polymer of four distinct monomers called *adenine* (A), *guanine* (G), *cytosine* (C), and *thymine* (T). DNA sequences are therefore naturally represented as words in the {A,C,G,T} alphabet, where the letters of the alphabet are called "nucleotides" or "bases". To estimate the number of substitutions in a

[1] Corresponding author's E-mail: silvia.lopes@ufrgs.br

DNA sequence, one needs a probabilistic model to describe changes between nucleotides. Continuous-time Markov chains are commonly used for this purpose. The nucleotide sites in the sequence are normally assumed to be evolving independently of each other. Substitutions at any particular site are described by a Markov chain, with the four nucleotides as being the *states* of the chain. The probability with which the chain jumps into other nucleotide states depends on the current state, but not on how the current state is reached. Besides this basic assumption, one often place further constraints on substitution rates between nucleotides, leading to different models of nucleotide substitution (see Yang, 2007). In this work we consider the nucleotide substitution models proposed by Jukes and Cantor (1969), Kimura (1980), Felsenstein and Churchill (1996) and Hasegawa et al. (1985)

The applications here mentioned, and many other important biological questions regarding evolutionary history and phylogenies, may be tackled by statistical analysis of DNA sequences using the likelihood function. In order to make a proper analysis, a base substitution model is needed. Since several of these models exist, statistical methods that choose in an appropriate way among them are important in this field (see Durbin et al., 2004; Felsenstein, 2004 and Yang, 2007).

The goal of this work is to study the likelihood ratio (LR) test properties for comparisons of base substitution models. Among all classes of statistical tests, the likelihood ratio test is the most widely used. Here we consider the hypothesis test of the form $H_0 : \boldsymbol{\theta}_r = \boldsymbol{\theta}_{r_0}$ versus $H_1 : \boldsymbol{\theta}_r \neq \boldsymbol{\theta}_{r_0}$, where $\boldsymbol{\theta} = (\boldsymbol{\theta}_r, \boldsymbol{\theta}_s) \in \Theta$ is the parameter vector of the substitution model under this hypotheses test.

There are two main reasons for using these tests to make sure the appropriate model is chosen. First of all, several procedures that use the likelihood function are model sensitive (see Goldman, 1993). Thus, the use of an inappropriate model might affect the analysis. On the other hand, some of the available base substitution models have many parameters, making the whole analysis computationally intensive. Thus, one wants to select a complex model only if it presents a significant increase in performance, compared to simpler versions. Another reason for investing time in model selection is that these models often represent specific biological features of the DNA evolving process. Thus, deciding which is the best model can bring insights to the actual evolution of a particular gene. For either reason, biologists frequently use LR tests for model selection, and they have been implemented in a number of bio-informatics softwares, such as PALM (see Yang, 2007) and MODELTEST (see Posada and Crandall, 1998). These tests are usually concerned with the probability of selecting a complex model by simple chance, when in fact the simple model is correct (type I error), but they seldom consider the probability that the null hypothesis is not rejected due to the lack of statistical power, and not to biological reasons. To address this matter, we study here the power of these LR tests for base substitution model selection.

Through asymptotic theory, we propose a low computational cost estimator for the power of the likelihood ratio test. This easy implementation and quick response estimator is based both on the asymptotic non-central $\chi^2_k(D)$ distribution of the LR statistic test, under H_1, and on a consistent estimator for the non-centrality parameter D. We also study the distribution of the test's statistic through Monte Carlo simulations. As an application of the methodology, we consider 17 different sequences of the ECP-EDN primate gene family.

The paper is organized as follows. In Section 2, we present the basic definitions for the models of DNA base substitution. In Section 3 we present the assumptions under which the likelihood ratio tests are built. The regularity conditions and the statistic of the test are also presented there. Section 4 states the main results of this work related to the power and the asymptotic normal distribution of the LR test, under H_1, when the sequence length goes to infinity. The proof of all results are given in the Appendix. Section 5 presents the estimation of the non-centrality parameter D and, as a consequence, we also present an estimator for the power of the LR test that requires a low computational effort. In Section 6 a Monte Carlo simulation study is conducted together with the analysis of 17 primate DNA sequences. Section 7 concludes the paper.

2 Models of DNA Base Substitution

Continuous time Markov chains are commonly used to assign probabilities to mutational events. The commonly used models assume that the nucleotide sites of the DNA sequence evolve independently, according to the same process (see Durbin et al, 2004 and all references therein). Base substitutions at any given site are described by a Markov chain with the four nucleotides (Adenine A, Guanine G, Cytosine C and Thymine T) as its states. The process is assumed to have reached stationarity.

For the sake of simplicity, we shall denote in the sequel the four distinct nucleotides by the numbers 1 to 4. The rate matrix of substitution models $\mathbf{Q}$ is given by

$$\mathbf{Q} = \begin{pmatrix} h_1 & \alpha_2\pi_2 & \alpha_4\pi_3 & \alpha_6\pi_4 \\ \alpha_1\pi_1 & h_2 & \alpha_8\pi_3 & \alpha_{10}\pi_4 \\ \alpha_3\pi_1 & \alpha_7\pi_2 & h_3 & \alpha_{12}\pi_4 \\ \alpha_5\pi_1 & \alpha_9\pi_2 & \alpha_{11}\pi_3 & h_4 \end{pmatrix}, \tag{2.1}$$

where the value h_l is such that the sum of elements in row l is 0, for all $l \in \{1, \cdots, 4\}$. Note that the stationary vector for this process is

$$\mathbf{p}_0 = (\pi_1, \pi_2, \pi_3, \pi_4), \tag{2.2}$$

where π_i represents the proportion of base i in the sequences, with $i \in E = \{1, 2, 3, 4\}$.

Four different models for this process are considered in this paper: the model JC69 (Jukes and Cantor, 1969) assumes that all bases have the same frequencies in the sequences and that all mutations have equal probabilities; the model K80 (Kimura, 1980) also assumes the homogeneous distribution for the base frequencies, but assesses different probabilities for transitions and transversions; the model F81 (Felsenstein and Churchill, 1996) allows different probabilities for the bases and assumes that mutations are proportional to base frequencies; and the model HKY85 (Hasegawa et al., 1985) allows different probabilities for the bases, but distinguishes between transitions and transversions. The constraints imposed by these models on the rates of matrix (2.1) are presented in Table 2.1.

The likelihood function $\mathcal{L} \equiv \mathcal{L}(\mathrm{M}|\mathbf{X})$ is defined as the probability of a certain model M, given the observed data $\mathbf{X}$ and a known topology F. It can be obtained as the probability that the model M assigns to the data $\mathbf{X}$, regarding that the phylogenetic structure related to the sequences is observed. For each site, given the phylogeny, this probability is obtained by adding the probability of mutational events, over all possible base combinations at the

Table 2.1: Constraints on Model Parameters.

Model	$\mathbf{p}_0$	Mutation Rates
JC69	$(\frac{1}{4},\frac{1}{4},\frac{1}{4},\frac{1}{4})$	$\alpha_1=\alpha_2=\alpha_3=\alpha_4=\alpha_5=\alpha_6=$ $\alpha_7=\alpha_8=\alpha_9=\alpha_{10}=\alpha_{11}=\alpha_{12}$
K80	$(\frac{1}{4},\frac{1}{4},\frac{1}{4},\frac{1}{4})$	$\alpha_1=\alpha_2=\alpha_{11}=\alpha_{12}$; $\alpha_3=\alpha_4=\alpha_5=\alpha_6=\alpha_7=\alpha_8=\alpha_9=\alpha_{10}$
F81	$(\pi_1,\pi_2,\pi_3,\pi_4)$	$\alpha_1=\alpha_2=\alpha_3=\alpha_4=\alpha_5=\alpha_6=$ $\alpha_7=\alpha_8=\alpha_9=\alpha_{10}=\alpha_{11}=\alpha_{12}$
HKY85	$(\pi_1,\pi_2,\pi_3,\pi_4)$	$\alpha_1=\alpha_2=\alpha_{11}=\alpha_{12}$; $\alpha_3=\alpha_4=\alpha_5=\alpha_6=\alpha_7=\alpha_8=\alpha_9=\alpha_{10}$

internal nodes of the tree. Since independent sites are assumed, the likelihood function of the entire sequence is obtained as the product of that probability for each site. Felsenstein (1981) proposes the use of a pruning algorithm to make these computations feasible. Although neighboring nucleotides in a DNA sequence are not always independent, the models do assume independence of evolution at different sites.

3 Likelihood Ratio Tests

The likelihood ratio (LR) test is one of the most used method for choosing between substitution models. All models presented here have the following common assumptions:

A1. The sequences are related by a phylogeny;

A2. Mutations in all sites are independent and identically distributed;

A3. Mutation probabilities are given by a Markov chain with rate matrix in the form of expression (2.1).

The distinction between models is made by the different constraints on matrix $\mathbf{Q}$. Thus, let

$$\begin{aligned} H_0 &: \mathbf{Q} = \mathbf{Q}_{\mathrm{M}_0} \\ H_1 &: \mathbf{Q} = \mathbf{Q}_{\mathrm{M}_1}, \end{aligned} \tag{3.1}$$

where M_0 and M_1 represent base substitution models from Table 2.1.

The statistic of the likelihood ratio test for comparing H_0 versus H_1 is given by

$$-2\Delta(\mathbf{X}) = -2\left(\log\left(\hat{\mathcal{L}}(\mathrm{M}_0|\mathbf{X})\right) - \log\left(\hat{\mathcal{L}}(\mathrm{M}_1|\mathbf{X})\right)\right), \tag{3.2}$$

where $\hat{\mathcal{L}}(\mathrm{M_j}|\mathbf{X})$ is the maximum of the likelihood function under H_j, for $j \in \{0,1\}$.

The asymptotic behavior of $-2\Delta(\mathbf{X})$ can be used in order to assess the test significance level. Let $F_{\mathbf{X}}(\cdot)$ be the distribution function of the random matrix $\mathbf{X}$ and $\boldsymbol{\theta}$ be its parameter vector. Suppose the following regularity conditions

$$\left\|\mathbb{E}\left(\frac{\partial}{\partial\theta_u}\log(\mathcal{L})\frac{\partial}{\partial\theta_w}\log(\mathcal{L})\right)\right\| < \infty \tag{3.3}$$

and

$$\mathbb{E}\left(\frac{\partial}{\partial\theta_u}\log(\mathcal{L})\frac{\partial}{\partial\theta_w}\log(\mathcal{L})\right)+\mathbb{E}\left(\frac{\partial^2}{\partial\theta_u\partial\theta_w}\log(\mathcal{L})\right)=\frac{\partial^2}{\partial\theta_u\partial\theta_w}\int\mathcal{L}=0, \tag{3.4}$$

hold, for all $\theta_u, \theta_w \in \boldsymbol{\theta}$.

Consider the hypotheses test

$$H_0 : \boldsymbol{\theta}_r = \boldsymbol{\theta}_{r_0} \quad \text{versus} \quad H_1 : \boldsymbol{\theta}_r \neq \boldsymbol{\theta}_{r_0}. \tag{3.5}$$

Then, under H_0, the statistic $-2\Delta(\mathbf{X})$ of the likelihood ratio test converges in distribution to a Chi-square χ^2_k distribution, with k degrees of freedom, as the sequence length S goes to infinity. Notice that the degrees of freedom k are determined by the difference of the free parameters between the two models which are being tested (see Wilks, 1962).

For the models in Table 2.1, the regularity conditions (3.3) and (3.4) are satisfied, as long as the tree topology is known (see Cybis, 2009).

There are many studies in the literature that assess the asymptotic distribution of the test statistic in expression (3.2), under H_0. See, for instance, Goldman (1993) and Whelan and Goldman (1999). These authors have shown, by simulation, that the χ^2_k distribution is achieved for small values of S, such as $S = 50$ or 100.

4 Test Power

The reason for using a likelihood ratio test to choose between any two substitution models is the selection of a more complex model only if it represents a significant improvement in performance. Thus, the null hypothesis is rejected only if the probability that the model M_0 generates data $\mathbf{X}$ is very small. On the other hand, when a simpler model is not rejected, the probability that the alternative model M_1 is better for the data, but the test was unable to detect that, is rarely considered.

The power of the test, defined as the probability that the test rejects H_0, given that H_1 is true, is an important feature for model selection. Thereafter, it should be always considered, whenever an informed decision on the best substitution model is desired.

An asymptotic result by Wald (1943) is available to determine the power of the test. Let $\boldsymbol{\theta}_r$ be the parameter vector of the substitution model under the hypotheses test in (3.5). Then, assuming that the regularity conditions (3.3) and (3.4) hold, the asymptotic distribution of $-2\Delta(\mathbf{X})$, under H_1, is a non-central $\chi^2_k(D)$ distribution, with k degrees of freedom and non-centrality parameter D. The non-centrality parameter D depends on S and is given by

$$D_S = (\boldsymbol{\theta}_r - \boldsymbol{\theta}_{r_0})\mathbf{I}_S(\boldsymbol{\theta}_r - \boldsymbol{\theta}_{r_0})', \tag{4.1}$$

where $\mathbf{I}_S$ is the *Fisher information matrix* defined as

$$\mathbf{I}_S \equiv \left(\mathbb{E}\left[\frac{\partial^2\log(\mathcal{L}(\boldsymbol{\theta}|\mathbf{X}_1,\cdots,\mathbf{X}_S))}{\partial\theta_u\partial\theta_w}\right]\right)_{\theta_u,\theta_w\in\boldsymbol{\theta}_r} = (I_{u,w})_{u,w}. \tag{4.2}$$

Notice that, for the models in Table 2.1, this result also holds, since it relies on the same regularity conditions needed to establish the convergence under H_0.

Thus, the power of the LR test can be obtained as

$$\widehat{\text{Power}} = \mathbb{P}\left(\chi_k^2(\widehat{D}_S) > \chi_{k,\alpha}^2\right), \tag{4.3}$$

where $\chi_{k,\alpha}^2$ represents the α%-quantile of the central χ_k^2 distribution.

The above result states that for large sequence length the distribution of $-2\Delta(\mathbf{X})$, under H_1, approaches a non-central Chi-square distribution. However, if we use even longer sequences, then we have the result given by the following theorem.

Theorem 4.1. *Let the likelihood ratio test, given in expression* (3.5), *be the test for comparing two substitution models. Let the statistic* $-2\Delta(\mathbf{X})$ *of the LR test be given by the expression* (3.2). *Then, the asymptotic distribution of* $-2\Delta(\mathbf{X})$, *under* H_1, *is given by*

$$\frac{-2\Delta(\mathbf{X}) - (k + D_S)}{\sqrt{2k + 4D_S}} \xrightarrow{d} \mathcal{Z}, \quad \text{when } S \to \infty, \tag{4.4}$$

where the degrees of freedom k *are determined by the difference of the free parameters between the models under* H_1 *and* H_0*, respectively,* D_S *is defined by* (4.1) *and* $\mathcal{Z} \sim N(0,1)$.

Proof: See the Appendix. □

While the asymptotic distribution of the test statistic under the null hypothesis depends only on the number of free parameters of the models, under the alternative hypothesis it depends on a number of different factors. Consequently, the power of the LR tests cannot be determined beforehand as it is done for the critical value.

Remark 4.1. One of the main factors that affect the power of the test is the sequence length S. It can be easily shown that the power of the LR test approaches 1 as S increases. Under i.i.d. site assumption, it is easy to show that $I_S = SI_1$ (see Cybis, 2009), where

$$I_1 \equiv \left(-\mathbb{E}\left[\frac{\partial^2 \log(\mathcal{L}(\boldsymbol{\theta}|\mathbf{X}_1))}{\partial\theta_u\partial\theta_w}\right]\right)_{\theta_u,\theta_w\in\boldsymbol{\theta}_r}.$$

Therefore, the expression (4.1) can be rewritten as

$$D_S = (\boldsymbol{\theta}_r - \boldsymbol{\theta}_{r_0})\mathbf{I}_S(\boldsymbol{\theta}_r - \boldsymbol{\theta}_{r_0})' = S(\boldsymbol{\theta}_r - \boldsymbol{\theta}_{r_0})\mathbf{I}_1(\boldsymbol{\theta}_r - \boldsymbol{\theta}_{r_0})'. \tag{4.5}$$

From the expression (4.5) above, when $S \to \infty$, the non-centrality parameter D_S also goes to infinity. And $\mathbb{P}(\chi_k^2(D_S) > \chi_{k,\alpha}^2) \to 1$, when $D_S \to \infty$.

The power of the LR test is also affected by the values of the parameters being tested. All other factors remaining constant, it is quite straightforward to observe that the greater the difference among the actual values of the parameters and the ones under H_0, the greater the power of the LR test. These both results are expected. The sequence length S represents here the sample size, and usually longer sequences result in an increase of the ability to distinguish between any two models. And the difference among the true parameter values and the ones specified under H_0 is what we are trying to decide about.

5 Estimation of the Power of the Test

We observe that the non-centrality parameter D_S, given in (4.1), depends on the Fisher information matrix, which does not have a closed expression for these models (because of the different trees being used). So, generally, there is no closed form expression for the parameter D_S. Nevertheless, the power of the LR test can be estimated. In this section we shall present a computational and a theoretical approach for its estimation.

5.1 Monte Carlo Simulations

Monte Carlo simulations have been used to assess different characteristics of the distribution of $-2\Delta(\mathbf{X})$, under H_0. See, for instance, Goldman (1993), Whelan and Goldman (1999) and Yang et al. (1994). Here we use a similar approach to study the test power. For this purpose, we simulate sequences according to the known phylogenetic tree, obtain the test statistic and accumulate the results, under both models of H_0 and H_1. A positive aspect of Monte Carlo simulations is that, unlike theoretical asymptotic distributions, it doesn't depend on large sample size to be accurate. It can be used to assess the exact distribution of the LR test statistic for any sequence length. Due to the many replications needed to obtain reliable information of the distributions being studied, Monte Carlo simulations have high computational cost. And this cost only increases with the complexity of the models. In many situations these simulations are not practical to assess the power of the test in every day applications. In view of this, in the next section, we discuss an easier procedure to obtain the power.

5.2 Estimation of the Non-centrality Parameter

The power estimation of the LR test by Monte Carlo simulations involves considerable computational effort. Thus, it is a suitable method for studying the test properties, but not practical enough to be considered whenever the LR test is applied to biological data. With the aid of the asymptotic theory, we propose an estimator for the power of the LR test that does not require more computational effort than already needed for obtaining the test statistic.

Under the alternative hypothesis H_1, the asymptotic distribution of $-2\Delta(\mathbf{X})$ is a non-central $\chi^2_k(D_S)$ distribution. The degrees of freedom k are easily determined by the difference of free parameters between the two models under H_1 and H_0, respectively. However, the non-centrality parameter D_S, defined in expression (4.1), depends on the actual values of the parameters under testing, and on the Fisher information matrix given by (4.2). In real cases, since these quantities are unknown, they must be estimated. To estimate the Fisher information matrix, we use the *observed Fisher information matrix* $\hat{\mathbf{I}}_S = (\hat{I}_{u,w})_{u,w}$, whose entries are given by

$$\hat{I}_{u,w} = - \left. \frac{\partial^2 \log(\mathcal{L})}{\partial\theta_u \partial\theta_w} \right|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} \tag{5.1}$$

for all $\theta_u, \theta_w \in \boldsymbol{\theta}_r$, where $\mathcal{L}$ is the likelihood function and $\widehat{\boldsymbol{\theta}}$ is the maximum likelihood estimator for $\boldsymbol{\theta}$.

Most computational packages that consider the likelihood function for the analysis of DNA sequences already use the observed Fisher information matrix for estimating the variance of the maximum likelihood estimates. Thus, power estimation through this statistic would not represent an increase in computational effort.

In the following lemma we propose an estimator for the non-centrality parameter D_S and for the LR test power.

Lemma 5.1. *Let the statistic* $-2\Delta(\mathbf{X})$ *of the* LR *test be given by expression* (3.2). *Consider its asymptotic distribution, under* H_1, *given by Theorem 4.1. Then, the non-centrality parameter* D_S, *given in* (4.1), *can be estimated by*

$$\widehat{D}_S = (\widehat{\boldsymbol{\theta}}_r - \boldsymbol{\theta}_{r_0})\hat{\mathbf{I}}_S(\widehat{\boldsymbol{\theta}}_r - \boldsymbol{\theta}_{r_0})', \tag{5.2}$$

where $\widehat{\boldsymbol{\theta}}_r$ *is the maximum likelihood estimator for* $\boldsymbol{\theta}_r$ *and* $\hat{\mathbf{I}}_S$ *is the observed Fisher information matrix, given in expression* (5.1).

Moreover, the test power can be estimated by

$$\widehat{Poder} = \mathbb{P}\left(\chi^2_k(\widehat{D}_S) > \chi^2_{k,\alpha}\right), \tag{5.3}$$

where the degrees of freedom k *are determined by the difference of the free parameters between the models under* H_1 *and* H_0, *respectively,* $\chi^2_k(\widehat{D}_S)$ *is a non-central Chi-square distribution, with* k *degrees of freedom and non-centrality parameter* $\widehat{D}_S$ *and* $\chi^2_{k,\alpha}$ *represents the* $\alpha\%$*-quantile of the central* χ^2_k *distribution.*

Proof: The proof follows immediately from the definitions of the non-centrality parameter D_S and of the LR test power. □

The following theorem establishes the consistency of the estimator $\widehat{D}_S$, proposed in Lemma 5.1. Note that an estimator $\hat{\theta}$ is *consistent* for the parameter θ if and only if $\hat{\theta} \xrightarrow{p} \theta$ (that is, $\lim_{n\to\infty}\mathbb{P}(|\hat{\theta}-\theta|<\epsilon)=1$, for all $\epsilon>0$).

Theorem 5.1. *Let the likelihood ratio test, given in expression* (3.1), *be the test for comparing two substitution models. Let* $\widehat{D}_S$ *be the estimator for the non-centrality parameter* D_S *be given by Lemma 5.1. Then,* $\widehat{D}_S$ *is a consistent estimator for the parameter* D.

Proof: See the Appendix. □

6 Simulations and Analysis of DNA Sequences

In this section we present a Monte Carlo simulation study to analyze the result of Theorem 4.1 and an application based on 17 different gene sequences of the ECP-EDN primate family.

6.1 Monte Carlo Simulations

Here we present a Monte Carlo simulation study to see how different sequence lengths S, different number N of species and different parameters for the models can affect the results in Theorem 4.1. For this purpose, we consider three different trees: Tree 1 has $N = 4$; Tree 2 has $N = 13$ and Tree 3 has $N = 19$ different species (see Figure 6.1). We also consider four LR tests for comparing 4 different base substitution models based on these trees: JC69 × K80, JC69 × F81, K80 × HKY85 and F81 × HKY85.

These three phylogenetic trees, originally estimated from real DNA sequences, are used as examples in the PAML package (see Yang, 2007). For each phylogenetic tree, descendent sequences were generated with length $S \in \{100; 500; 1,000; 2,000\}$, according to the four models involved in hypotheses test. For the transition and transversion ratio rate $K = \alpha_1/\alpha_3$ of the base substitution models, from Table 2.1, we consider $K \in \{3, 5\}$ for the model K80 and $K = 3$ for the model HKY85. For models F81 and HKY85 we consider $\mathbf{p}_0 = (0.2, 0.3, 0.3, 0.2)$ while, for models JC69 and K80, the values of $\mathbf{p}_0$ are the ones given in Table 2.1. For all simulations we consider $Re = 1,000$ replications. The choice of the parameter values was based on real sequence analysis so that they would make biological sense.

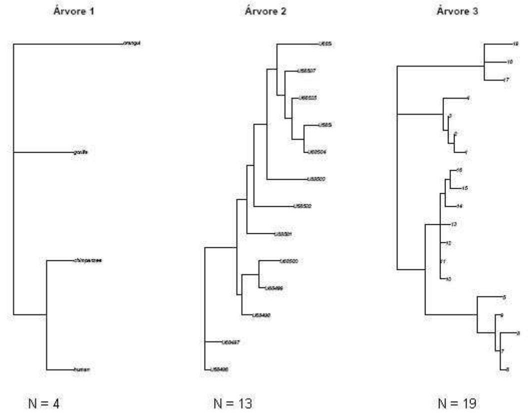

Figure 6.1: Simulation Trees.

Figure 6.2 shows the estimated distribution for the test statistic $-2\Delta(\cdot)$, under H_1, for the JC69 × K80 test, based on Tree 1, for different sequence lengths S. One observes that for sequence length $S = 500$, the shape of the distribution is already very close to the normal distribution.

To assess the distribution of the statistic $-2\,\Delta(\cdot)$, given by the expression (3.2), we consider the Shapiro-Wilks goodness of fit test (see Wilks, 1962) for each pair of four models based on each tree in Figure 6.1. The results are reported in Table 6.1. From Table 6.1 we observe that the statistic $-2\,\Delta(\mathbf{X})$ has normal distribution whenever the sequence length $S \in \{6,000; 10,000\}$, for all four tests considered here. For this Monte Carlo experiment, when $S = 100$, the Shapiro-Wilks test rejects the hypothesis that the distribution of the statistic $-2\,\Delta(\mathbf{X})$ is normally distributed, for any of the four models. When $S = 2,000$, only for models JC69 × K80, considering any tree and for F81 ×

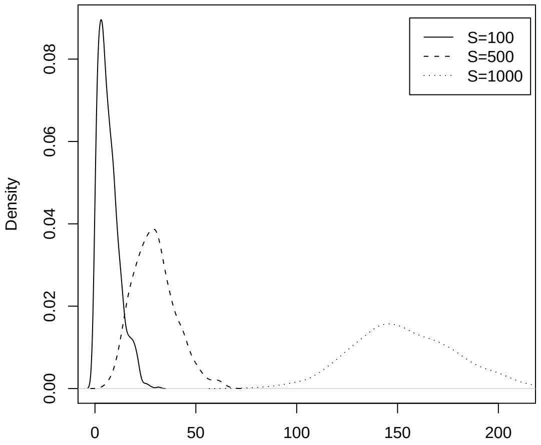

Figure 6.2: Estimated Distribution of $-2\Delta(\cdot)$, under H_1, for JC69 × K80 Test, Based on Tree 1, for Different Sequence Lengths S.

HKY85, when using Tree 2, the Shapiro-Wilks test accepts the normal distribution for the statistic, under the null hypothesis of normality. Therefore, from Table 6.1, we observe that Theorem 4.1 holds for sequence length of at least $6,000$ for the considered tests and trees.

No direct correlation between the number of sequences N and the power of the test were noted in the simulations. Tree 2, with intermediate number of sequences $N = 13$, presented the highest test power for all conducted simulations. But further investigations (see Cybis et al., 2010) show strong correlation between test power and total tree length.

Table 6.1: Shapiro-Wilks Test For Testing Normality.

Test	Tree	$S = 100$		$S = 2,000$		$S = 6,000$		$S = 10,000$	
		Statistic	p-value	Statistic	p-value	Statistic	p-value	Statistic	p-value
JC69 × K80	1	0.9592	4.427e-16*	0.9977	0.1733	0.9981	0.3066	0.9987	0.6856
	2	0.9863	4.727e-08*	0.9986	0.6550	0.9967	0.0379	0.9984	0.4755
	3	0.9662	1.749e-14*	0.9974	0.1073	0.9984	0.5029	0.9979	0.2833
F81 × HKY85	1	0.9640	5.353e-15*	0.9958	0.0084*	0.9973	0.0899	0.9991	0.9009
	2	0.9830	2.167e-09*	0.9990	0.8517	0.9973	0.0894	0.9980	0.2933
	3	0.9714	3.936e-13*	0.9958	0.0085*	0.9968	0.0469	0.9983	0.4459
JC69 × F81	1	0.9584	2.997e-16*	0.9942	0.0007*	0.9961	0.0142	0.9977	0.2091
	2	0.9629	2.932e-15*	0.9949	0.0020*	0.9988	0.7381	0.9984	0.5241
	3	0.9316	<2.20e-16*	0.9915	1.51e-05*	0.9980	0.2951	0.9983	0.4426
K80× HKY85	1	0.9552	<2.20e-16*	0.9932	0.0002*	0.9959	0.0103	0.9979	0.2769
	2	0.9672	3.130e-14*	0.9952	0.0031*	0.9982	0.3764	0.9962	0.0191
	3	0.9367	<2.20e-16*	0.9956	0.0060*	0.9983	0.4994	0.9993	0.9965

Note: All tests are conducted with 99% confidence level and symbol “ $*$ ” means rejection.

6.2 Analysis of DNA Sequences

We consider for the application 17 different sequences of the ECP-EDN primate gene family with $S = 483$. The ECP-EDN gene family is composed of two ribonucleases (ECP and EDN) that are responsible for the inespecific immune response in these animals. These sequences are also considered by Bielawski and Yang (2003). For the application analysis we consider the same phylogenic topology as obtained by Bielawski and Yang (2003). Figure 6.3 presents this phylogeny, where the circle near to the tree root represents a gene duplication event.

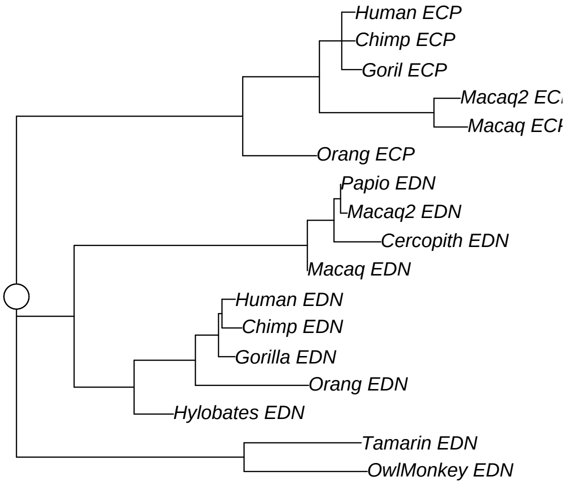

Figure 6.3: Phylogenetic Tree for the Primate Gene Family ECP-EDN, with $S = 483$.

As in the Monte Carlo simulations (see Section 6.1), here we also consider four models: JC69, K80, F81 and HKY85. The maximum likelihood estimates for the parameters of the last three models are presented in Table 6.2. The software PAML, used in this analysis, sets all α parameters of the JC69 model to 1. In Table 6.2 we present the estimates of the model parameters, with their asymptotic standard error (denoted by ase), given by the PAML software. To give also the bias, standard error (se) and mean square error (mse) values of these estimates, we consider the Jackknife methodology (see Efron, 1982) to produce 483 different samples (one for each site of the sequence) for each model. One observe, from Table 6.2, that for each estimate, the standard error obtained from the Jackknife procedure is very similar to the asymptotic standard error. The mean square error values are small for all nucleotide probabilities. We also consider four model tests: JC69 × K80, JC69 × F81, K80 × HKY85 and F81 × HKY85. The results for these four tests are presented in Table 6.3.

The distributions of the test statistic $-2\Delta(\cdot)$, both under H_0 and under H_1, were obtained by Monte Carlo simulation. Figure 6.4 presents the histograms of the $-2\Delta(\cdot)$ distribution, both under H_0 and H_1, for all four tests. The vertical dotted line for the histograms, under H_0, represents the observed test statistic value evaluated for the data $\mathbf{X}$. Under H_1, the vertical dotted line represents the critical value of 99% confidence level obtained from the Monte Carlo simulation.

The test power is also obtained using the proposed estimator $\widehat{D}_S$, from Lemma 5.1 for all four tests. We observe that the values of the test power obtained from this method are very close to the ones obtained from the Monte Carlo simulations (see the asymptotic

Table 6.2: Maximum Likelihood Estimates for each Model in the ECP-EDN Primate Gene Family.

Model	Estimates	ase	bias(J)	se(J)	mse(J)
K80	$\widehat{K} = 2.09392$	0.26919	-0.03051	0.29125	0.08576
F81	$\hat{p}_1 = 0.28002$	0.01678	-0.00181	0.01686	0.00029
	$\hat{p}_3 = 0.25721$	0.01627	0.00529	0.01667	0.00031
	$\hat{p}_4 = 0.25478$	0.01619	-0.00359	0.01785	0.00033
HKY85	$\widehat{K} = 2.12470$	0.27429	0.03376	0.29796	0.08992
	$\hat{p}_1 = 0.28615$	0.01709	0.00065	0.01710	0.00029
	$\hat{p}_3 = 0.25401$	0.01619	0.00559	0.01677	0.00031
	$\hat{p}_4 = 0.25257$	0.01611	-0.00107	0.01778	0.00032

Note: The estimate $\hat{p}_2$ is obtained by differencing the estimates $\hat{p}_i$, for $i \in \{1,3,4\}$, from 1.
The symbol "(J)" means the statistic is obtained from the Jackknife procedure.

distribution and Monte Carlo simulation in Table 6.3). The results of Table 6.3, which are corroborated by a similar and more detailed analysis in Cybis et al. (2010), indicate the good performance of estimator $\widehat{D}_S$. If we add this to its low computational cost, we have that test power estimation through $\widehat{D}_S$ is a good candidate for implementation in bio-informatic softwares which consider LR tests.

We observe that for JC69 × K80 and F81 × HKY85 tests, where we are testing different rates for *transitions* and *transversions*, the null hypothesis is always rejected. Therefore, we conclude that different transitions and transversion rates played an important role for the evolution of this gene family. The power value for both tests are very close to 1.

However, for tests that compare models with or without the base frequency homogeneity assumption, such as JC69 × F81 and K80 × HKY85, the null hypothesis is accepted for both tests. Hence, we observe that the base frequencies for this genic family are sufficiently close to the uniform distribution. Therefore, it seems that the use of three additional parameters representing the base frequencies in further likelihood function analysis is not justified. And, for parsimony reasons one should prefer a simpler model. However, from Table 6.3 we observe that the power for both tests are very close to 50%. Therefore, there exists a considerable probability that the alternative hypothesis is better and the test doesn't have sufficient power to reject the null hypothesis. So, this result must be used with caution.

When comparing the four models JC69, K80, F81 and HKY85, based on these tests, we suggest the use of the K80 model for better describing the evolution process of these sequences, since this model allows for different transition and transversion rates in a homogeneous base frequency set up.

In the literature, the most largely used model is HKY85, followed by simpler models (see Bielawski and Yang, 2003). The reason behind this is that it is one of the simplest models that allow for both transition and transversion bias and inhomogeneous base frequencies, two important biological features. Note that, more complex models, with more parameters, represent a considerable increase in computational effort.

It is also noteworthy that the i.i.d. sites hypothesis, used in this work, is very restricting and may be biologically inaccurate. The impact of different assumptions for the inter sites distribution in these tests, particularly in the framework of the test power, is analyzed in further work (see Cybis et al. 2010) to be published elsewhere.

Table 6.3: Test Results for the Primate Family.

Test	$-2\Delta(\mathbf{X})$	Asymptotic Distribution			Monte Carlo Simulation			Decision
		k	p-value	Power	p-value	Critical Value 99%	Power	
JC69 × K80	32.0992	1	1.4650×10^{-8}	0.9343	< 0.0010	6.7860	1	reject
F81 × HKY85	33.1716	1	8.4374×10^{-9}	0.9412	< 0.0010	6.2904	0.9990	reject
JC69 × F81	8.2774	3	4.0613×10^{-2}	0.4707	0.0440	11.6883	0.4680	accept
K80 × HKY85	9.3498	3	2.4984×10^{-2}	0.5340	0.0270	11.6545	0.5360	accept

7 Conclusions

In this paper we consider the likelihood ratio tests for comparing base substitution models. With mild regularity conditions, we present some of its properties such as the fact that the asymptotic distribution of the likelihood ratio test statistic is a non-central $\chi^2_k(D_S)$ distribution, under the alternative hypothesis H_1. The degrees of freedom k are determined by the difference of the free parameters between both models, under H_1 and H_0, respectively.

We also prove the asymptotic normal distribution of the likelihood ratio test statistic when the sequence length S goes to infinity. We present a Monte Carlo simulation study to see how different sequence lengths S, different number N of species and different parameters for the models could affect this result. We reject the normality hypothesis of the statistic $-2\,\Delta(\cdot)$ distribution, when $S = 100$, for any of the four models considered here and for most tests when $S = 200$. We do not reject the normality hypothesis when the sequence length S is at least $6,000$, for the considered tests and trees.

We propose a consistent estimator for the non-centrality parameter D_S which is easy to implement and has very quick response for these base substitution models. Through asymptotic theory and based on this consistent estimator for D_S, we propose a low computational cost estimator for the power of the likelihood ratio test.

We apply the methodology presented here to 17 different gene sequences of the ECP-EDN family in primates with $S = 483$. These sequences are also considered by Bielawski and Yang (2003). To evaluate the performance of the proposed test power estimator, we compare its estimates with Monte Carlo simulations for this set of data. It seems that the different base substitution models that are being tested and the sample characteristics have a great influence on both estimates. Nevertheless, the power of these tests estimated by both methods present very similar values. We observe that for JC69 × K80 and F81 × HKY85 tests, where we are testing different rates for *transitions* and *transversions*, the null hypothesis is always rejected. Therefore, we conclude that different transition and transversion rates played an important role for the evolution of these genes in the primate family. The power value for both tests are very close to 1.

Acknowledgments

G.B. Cybis was supported by CNPq-Brazil. S.R.C. Lopes research was partially supported by CNPq-Brazil, by CAPES-Brazil, by CNPq-INCT em Matemática and also by Pronex

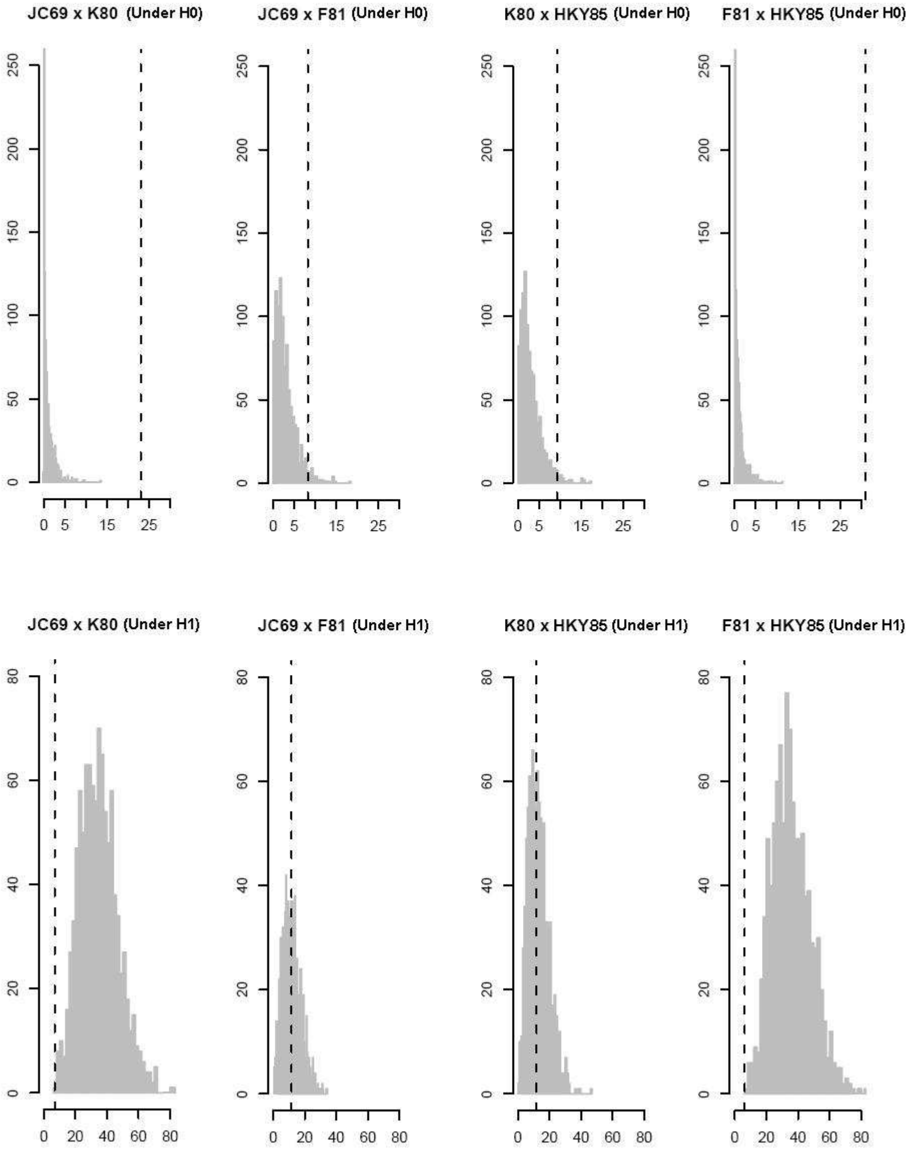

Figure 6.4: Histograms of $-2\Delta(\mathbf{X})$, under H_0 and H_1, for the Tests JC69 × K80, JC69 × F81, K80 × HKY85 and F81 × HKY85, under H_0 and H_1, for the Primate Sequences.

Probabilidade e Processos Estocásticos - E-26/170.008/2008 -APQ1. H.P. Pinheiro research was partially supported by CNPq-Brazil and by CAPES-Brazil.

The authors thank two anonymous referees and the editor for their valuable comments and suggestions that improved the final version of the manuscript.

References

Anisimova, S., Bielawski, J. and Yang, Z. (2001). "Accuracy and Power of the Likelihood Ratio Test in Detecting Adaptive Molecular Evolution". *Molecular Biology and Evolution*, Vol. **18**(8), 1585-1592.

Bielawski, P. and Yang, Z. (2003). "Maximum likelihood methods for detecting adaptive evolution after gene duplication". *Journal of Structural and Functional Genomics*, Vol. **3**, 201-212.

Cybis, G.B. (2009). *Teste da Razão de Verossimilhança e seu Poder em Árvores Filogenéticas.* Master Dissertation in the Graduate Program in Mathematics, Federal University of Rio Grande do Sul, Porto Alegre.

Cybis, G.B., Lopes, S.R.C. and Pinheiro, H.P. (2010). "Asymptotic Behavior of the Power of the Likelihood Ratio Test for Models of DNA Base Substitution". In preparation.

Durbin, R., Krogh, A., Eddy, S. e Mitchison, G. (2004). *Biological Sequence Analysis.* Cambridge: Cambridge University Press.

Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans.* Philadelphia: CBMS-NSF Regional Conference Series in Applied Mathematics.

Felsenstein, J. (2004). *Inferring Phylogenies.* Sunderland: Sinauer Associates.

Felsenstein, J. (1981). "Evolutionary Trees From DNA Sequences: A Maximum Likelihood Approach". *Journal of Molecular Evolution*, Vol. **17**, 368-376.

Felsenstein, J. and Churchill, G.A. (1996). "A Hidden Markov model approach to variation among sites in rate of evolution". *Molecular Biology and Evolution*, Vol. **13**, 93-104.

Goldman, N. (1993). "Statistical tests of models of DNA substitution". *Journal of Molecular Evolution*, Vol. **36**, 182-198.

Hasegawa, M., Kishino, H. and Yano, T. (1985). "Dating of human-ape splitting by a molecular clock of mitochondrial DNA". *Journal of Molecular Evolution*, Vol. **22**, 160-174.

Ho, S. and Phillips, M. (2009). "Accounting for Calibration Uncertainties in Phylogenetic Estimation of Evolutionary Divergence Times". *Systems Biology*, Vol. **58**(3), 367-380.

Johnson, N. and Kotz, S. (1970). *Continuous Univariate Distributions*, Vol. **2**. New York: Wiley.

Jukes, T. and Cantor, C. (1969). *Evolution of Protein Molecules.* New York: Academic Press.

Kimura, M. (1980). "A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences". *Journal of Molecular Evolution*, Vol. **16**, 111-120.

Posada, D. and Crandall, K.A. (1998). "MODELTEST: testing the model of DNA substitution". *Bioinformatics*, Vol. **14**, 817-818.

Shao, J. (2003). *Mathematical Statistics.* New York: Springer-Verlag.

Suzuki, Y. and Gojobori, T. (1999). "A method for detecting positive selection at single amino acid sites". *Molecular Biology and Evolution*, Vol. **16**(10), 1315-1328.

Wald, A. (1943). "Tests of Statistical Hypothesis Concerning Several Parameters when the Number of Observations is Large". *Trans. Amer. Math. Soc.*, Vol. **54**, 426-482.

Wilks, S. (1962). *Mathematical Statistics.* New York: Wiley.

Whelan, S. and Goldman, N. (1999). "Distributions of statistics used for the comparison of models of sequence evolution in phylogenetics". *Molecular Biology and Evolution*, Vol. **16**, 1292-1299.

Yang, M., Goldman, N. and Friday, A. (1994). "Comparison of Models for Nucleotide Substitution Used in Maximum-Likelihood Phylogenetic Estimation". *Molecular Biology and Evolution*, Vol. **11**(2), 316-324.

Yang, Z. (2007). "PAML 4: Phylogenetic Analysis by Maximum Likelihood". *Molecular Biology and Evolution*, Vol. **24**(8), 1586-1591.

Appendix

Proof of Theorem 4.1: Cybis (2009) showed that the regularity conditions given in expressions (3.3) and (3.4) are satisfied for all base substitution models considered here, as long as the phylogeny is known.

For any base substitution model and assuming independent sites, the likelihood function, based on a sample $\mathbf{X}$, is given by

$$\mathcal{L}(F, \bar{\tau}|\mathbf{X}) = \mathcal{L}(F, \bar{\tau}|X^1, \cdots, X^{2N-1}) = \prod_{u=1}^{S} \mathbb{P}(X_u^1, \cdots, X_u^{2N-1}|F, \bar{\tau}), \tag{A.1}$$

where F is the known tree topology, $\bar{\tau} = \{\tau_1, \tau_2, \cdots, \tau_{2N-2}\}$ represents branch lengths and $\mathbb{P}(X_u^1, \cdots, X_u^N|F, \bar{\tau})$ is the probability of generating N bases from F which is given by

$$\mathbb{P}(X_u^1, \cdots, X_u^N|F, \bar{\tau}) = \sum_{i^{N+1}, \cdots, i^{2N-1}} \pi_{i^{2N-1}} \prod_{k=N+1}^{2N-2} \mathbb{P}(i^k|i^{h(k)}, \tau_k) \prod_{l=1}^{N} \mathbb{P}(X_u^l|i^{h(l)}, \tau_l). \tag{A.2}$$

In expression (A.2), $\pi_{i^{2N-1}}$ is the frequency of the base that is in node $2N-1$ (phylogeny root). Hence, $\pi_{i^{2N-1}}$ represents the probability of finding the base i^{2N-1} in the u site of the sequence in the phylogeny root. The expression $\mathbb{P}(i^k|i^{h(k)}, \tau_k)$ represents the probability that the base in position u of the k node is generated from the base that is in the same position of its ancestral sequence $h(k)$ in τ_k time. This probability is denoted by $\mathbb{P}(X_u^l|i^{h(l)}, \tau_l)$ whenever the off-spring sequence belongs to the sample $\mathbf{X}$. In fact, in this case, one knows the X_u^l base. All these probability will depend on the base substitution model adopted.

In expression (A.1) there are 4^N possible base combinations for one position of all sequences $(X_1^u, \cdots, X_N^1)$. Let p_i be the probability of combination i and s_i be the number of times the combination i appears in the sample. Then,

$$\log\left(\mathcal{L}(F, \bar{\tau}|\mathbf{X})\right) = \sum_{i=1}^{N^4} s_i \log(p_i). \tag{A.3}$$

The function in expression (A.3) can be rewrite as a function of the statistics $\bar{s}_i = \frac{s_i}{S}$, for all $i \in \{1, \cdots, N^4\}$, such that $\bar{s}_i$ represents the site proportion in the sequence that has the combination i. Therefore,

$$\log\left(\mathcal{L}(F, \bar{\tau}|\mathbf{X})\right) = S \sum_{i=1}^{N^4} \bar{s}_i \log(p_i). \tag{A.4}$$

To calculate the non-centrality parameter D_S, given in expression (4.1), one needs the Fisher information matrix $\mathbf{I}_S = (I_{u,w})_{u,w}$, whose entries are given in expression (4.2). From the definition of the parameter D_S and from equation (A.4) one gets

$$\begin{aligned} D_S &= (\boldsymbol{\theta}_r - \boldsymbol{\theta}_{r_0})\mathbf{I}_S(\boldsymbol{\theta}_r - \boldsymbol{\theta}_{r_0})' \\ &= (\boldsymbol{\theta}_r - \boldsymbol{\theta}_{r_0}) \left[-S\mathbb{E}\left(\frac{\partial^2}{\partial\theta_u \partial\theta_w} \sum_{i=1}^{N^4} \bar{s}_i \log(p_i) \right) \right]_{\theta_u, \theta_w \in \boldsymbol{\theta}_r} (\boldsymbol{\theta}_r - \boldsymbol{\theta}_{r_0})' \\ &= S\left((\boldsymbol{\theta}_r - \boldsymbol{\theta}_{r_0}) \left[-\mathbb{E}\left(\frac{\partial^2}{\partial\theta_u \partial\theta_w} \sum_{i=1}^{N^4} \bar{s}_i \log(p_i) \right) \right]_{\theta_u, \theta_w \in \boldsymbol{\theta}_r} (\boldsymbol{\theta}_r - \boldsymbol{\theta}_{r_0})' \right). \end{aligned} \tag{A.5}$$

One observes that the probability p_i of the combination i is a function that depends on the phylogeny, the times $\bar{\tau}$ and the base substitution model parameters. However, this probability does not depend on S. Hence, the expression inside the parenthesis in equality (A.5) does not depend on S. Therefore, $\lim_{S\to\infty} D_S = \infty$.

From Johnson and Kotz (1970), if X is a random variable such that $X \sim \chi^2_k(D)$ with constant degrees of freedom k, then

$$\frac{X - (k + D_S)}{\sqrt{2k + 4D_S}} \xrightarrow{d} \mathcal{Z}, \quad \text{when } S \to \infty, \tag{A.6}$$

where $\mathcal{Z} \sim \mathbb{N}(0, 1)$ is the standard normal distribution with zero mean and variance equal to 1. The result in (4.4) follows immediately from (A.6). □

Proof of Theorem 5.1: One wants to show that $\widehat{D}_S \xrightarrow{p} D_S$. We observe that assuming the regularity assumptions (3.3) and (3.4), any maximum likelihood estimator is consistent (see Shao, 2003). From Cybis (2009), under i.i.d. sites assumption, all base substitution models considered here satisfy these regularity conditions. Then, $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}$. Also, from the i.i.d. site assumption, it is easy to see that

$$\begin{aligned} \mathbf{I}_S &= \left(-\mathbb{E}\left[\frac{\partial^2 \log(L(\boldsymbol{\theta}|\mathbf{X}_1, \cdots, \mathbf{X}_S))}{\partial\theta_u \partial\theta_w} \right] \right)_{\theta_u, \theta_w \in \boldsymbol{\theta}_r} = \left(-\mathbb{E}\left[\frac{\partial^2 \log\left(\prod_{i=1}^S L(\boldsymbol{\theta}|\mathbf{X}_i) \right)}{\partial\theta_u \partial\theta_w} \right] \right)_{\theta_u, \theta_w \in \boldsymbol{\theta}_r} \\ &= \left(-\sum_{i=1}^S \mathbb{E}\left[\frac{\partial^2 \log(L(\boldsymbol{\theta}|\mathbf{X}_i))}{\partial\theta_u \partial\theta_w} \right] \right)_{\theta_u, \theta_w \in \boldsymbol{\theta}_r} = S\mathbf{I}_1. \end{aligned} \tag{A.7}$$

Similarly,

$$\hat{\mathbf{I}}_S = \left(- \left. \frac{\partial^2 \log(L(\boldsymbol{\theta}|\mathbf{X}_1, \cdots, \mathbf{X}_S))}{\partial\theta_u \partial\theta_w} \right|_{\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}} \right)_{\theta_u, \theta_w \in \boldsymbol{\theta}_r} = \left(-\sum_{i=1}^S \left. \frac{\partial^2 \log(L(\boldsymbol{\theta}|\mathbf{X}_i))}{\partial\theta_u \partial\theta_w} \right|_{\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}} \right)_{\theta_u, \theta_w \in \boldsymbol{\theta}_r}.$$

From the weak law of large numbers (see Shao, 2003), one has

$$\frac{\hat{\mathbf{I}}_S}{S} = \frac{1}{S}\left(-\sum_{i=1}^{S} \frac{\partial^2 \log(L(\boldsymbol{\theta}|\mathbf{X}_i))}{\partial\theta_u\partial\theta_w}\bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}\right)_{\theta_u,\theta_w\in\boldsymbol{\theta}_r} \xrightarrow{p} \left(-\mathbb{E}\left[\frac{\partial^2 \log(L(\boldsymbol{\theta}|\mathbf{X}_i))}{\partial\theta_u\partial\theta_w}\bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}\right]\right)_{\theta_u,\theta_w\in\boldsymbol{\theta}_r}. \tag{A.8}$$

Notice that the expected value in expression (A.8) is a continuous function of $\hat{\boldsymbol{\theta}}$. Since $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}$, one can apply the Slutzky's theorem (see Shao, 2003). Hence, for each $i \in \{1, \cdots, S\}$, the right expression in (A.8) converges in probability to I_1. Therefore, the empirical Fisher information matrix is a consistent estimator for the Fisher information matrix, that is, $\hat{\mathbf{I}}_S \xrightarrow{p} S\mathbf{I}_1 = \mathbf{I}_S$. Since one has $\hat{\mathbf{I}}_S \xrightarrow{p} \mathbf{I}_S$ and $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}$, by Slutzky's theorem one has

$$\widehat{D}_S = (\hat{\boldsymbol{\theta}}_r - \boldsymbol{\theta}_{r0})\hat{\mathbf{I}}_S(\hat{\boldsymbol{\theta}}_r - \boldsymbol{\theta}_{r0}) \xrightarrow{p} (\boldsymbol{\theta}_r - \boldsymbol{\theta}_{r0})\mathbf{I}_S(\boldsymbol{\theta}_r - \boldsymbol{\theta}_{r0}) = D_S,$$

that is, $\widehat{D}_S$ is a consistent estimator for D_S. □