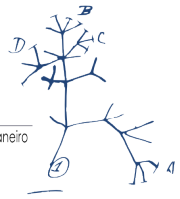
UFRJ

Programa de Pós-graduação em
Biodiversidade & Biologia Evolutiva
Instituto de Biologia - Universidade Federal do Rio de Janeiro



Lucas Pereira Marques

Testes topológicos e sinal conflitante entre marcadores filogenéticos: um estudo de caso com a raiz de Placentalia

Dissertação de Mestrado

Rio de Janeiro

Setembro de 2017

Lucas Pereira Marques

Testes topológicos e sinal conflitante entre marcadores filogenéticos: um estudo de caso com a raiz de Placentalia

Dissertação de mestrado apresentada ao Programa de Pós-Graduação em Biodiversidade e Biologia Evolutiva da Universidade Federal do Rio de Janeiro, para obtenção do título de Mestre em Biodiversidade e Biologia Evolutiva.

Orientadora: Profa. Dra. Claudia de Moraes Russo
Coorientador: Prof. Dr. Carlos Guerra Schrago

Rio de Janeiro

Setembro de 2017

Marques, Lucas

Testes topológicos e sinal conflitante entre marcadores filogenéticos: um estudo de caso com a raiz de Placentalia / Lucas Pereira Marques. – Rio de Janeiro, 2017

95 f.

Orientadora: Claudia de Moraes Russo
Coorientador: Carlos Guerra Schrago

Dissertação (Mestrado) – Universidade Federal do Rio de Janeiro, Instituto de Biologia, Programa de Pós-Grauação em Biodiversidade e Biologia Evolutiva, 2017.

1. Filogenômica 2. Testes topológicos 3. Heterogeneidade filogenética
I. Russo, Claudia. II. Schrago, Carlos. III. Universidade Federal do Rio de Janeiro.
IV. Dissertações

‘Liberdade é a liberdade de dizer que dois mais dois são quatro. Se isso for admitido, tudo o mais é decorrência.’

- Winston Smith

Agradecimentos

A minha orientadora por toda a paciência, apoio, referência profissional e, principalmente, por ter apostado no aluno de graduação que dormia tarde e acordava cedo demais.

Ao meu coorientador pela inestimável contribuição teórica a este trabalho e inspiração para os rumos que minha linha de pesquisa tomou.

Aos colegas de laboratório pela companhia e solicitude, especialmente a André Paschoa pelas longas reflexões sobre os temas da biologia evolutiva no período do mestrado e a Filipe Romero pelas tantas trocas de ideias que contribuíram para esta dissertação.

A Barbara Matias pela desenvoltura e paciência redobrada no árduo cargo de minha companheira ao longo do período mais difícil do mestrado.

Aos meus amigos da UFRJ e de toda a vida que continuaram me aceitando pelo que sou, apesar do desafio.

A minha família; quando o mundo cai, tudo o que tenho e tudo o que sou.

Resumo

Enquanto a humanidade testemunhava uma rápida evolução no poder de processammento dos computadores ao decorrer das últimas décadas, poucos suspeitariam que a disponibilidade de dados pudesse ultrapassar nossa capacidade de tratá-los de forma estatisticamente adequada. Hoje, particularmente no contexto da Filogenômica, teóricos ainda lutam para equilibrar modelos biologicamente realistas, viabilidade computacional e controle de erros em análises de conjuntos de dados genômicos. Mais especificamente no campo da verossimilhança, os meios mais comuns para avaliar a confiança de hipóteses filogenéticas divergentes são praticamente os mesmos testes topológicos já disponíveis desde 2002. Apesar de alguns desses testes demonstrarem acurácia que permanece indisputada (como o *Approximately Unbiased*), eles dependem de métodos de reamostragem, o que ameaça sua viabilidade para testar milhares ou mesmo milhões de árvores de genes. Aqui eu ensaio uma nova abordagem, independente de reamostragens, para testar topologias e comparo sua performance a de testes tradicionais. Em avaliações de múltiplas resoluções para a posição da raíz da árvore de Placentaria, o novo teste foi aproximadamente 10 vezes mais rápido que os testes KH, SH e AU com reamostragem acelerada por RELL. Todos os testes concordaram ao dar maior suporte, em geral, ao enraizamento de Placentalia entre os grupos Boreoeutheria e Atlantogenatha, mesmo essa resolução não tendo predominado entre as árvores de genes de máxima verossimilhança para os 747 loci nucleares analisados. Apesar do novo teste ser relativamente conservador e ainda experimental, após aprimoramentos e provas adicionais pode servir como um eficiente meio de prevenção ao uso de loci com sinal ambíguo, por exemplo em análises de coalescência que interpretam árvores de gene inferidas como observações.

Abstract

As mankind witnessed a rapid enhancement in computer processing power along the last decades, few would suspect that molecular data availability could surpass our capability to give it adequate statistical treatment. Today, particularly in the context of Phylogenomics, theoreticians still struggle to concur biologically realist models, computational tractability and error control on analyses of genome-sized datasets. More specifically under the likelihood framework, the standard means to assess confidence of differing tree hypotheses are roughly the same topological tests that were available in 2002. Although some of these tests display still-undisputed accuracy (like the Approximately Unbiased), they rely on resampling methods, which threats their feasibility for testing thousands or millions of gene trees. Here I try a novel, resample-independent approach to topology testing and compare its performance to that of traditional tests. On evaluations of multiple resolutions for the position of Placental tree root, the new test was approximately 10 times faster than KH, SH and AU tests with RELL accelerated resampling. All tests agreed in overall greater support to the rooting of Placentalia between Boroeutheria and Atlatogenatha cohorts, even though this resolution was not predominant among the maximum likelihood gene trees for the 747 nuclear loci analyzed. Although the new test is relatively conservative and still putative, after additional improvements and proofs it may serve, for instance, as an efficient fail-safe to coalescence methods that take inferred gene trees for observations, by selecting those genes with more unequivocal signal beforehand.

Lista de figuras

1: Curva de verossimilhança para a probabilidade de se obter cara (p) 2
2: Distribuição amostral de 1000 estimativas independentes de p 3
3: Precisão e acurácia das estimativas na presença de erro amostral vs erro sistemático 4
4: Intervalo de confiança do LRT dado pela função de χ^2 5
5: Matriz de probabilidades de transição (mudança) entre bases em função do tempo 6
6: Árvore filogenética e função de verossimilhança correspondente 8
7: Variação da topologia e da verossimilhança da árvore conforme mudam seus ramos internos. 9
8: Bootstrap de Felsenstein (1985) 11
9: Intervalo de confiança (IC) do teste KH pressupondo a normalidade da distribuição de $\Delta\ln L$. 13
10: Cópias ortólogas e parálogas ao longo da história evolutiva. 17
11: Coalescência de linhagens gênicas ao longo da árvore de espécies e as árvores de gene resultantes 19
12: Três resoluções mais aceitas para o posicionamento da raiz de Placentalia 22
13: Geração das árvores e parâmetros de substituição que guiaram as simulações dos dados. 27
14: Rearranjo por NNI de uma árvore com 16 táxons 28
15: Representação com árvores não enraizadas, de 8 táxons, do esquema de rearranjos realizado com cada árvore verdadeira 30
16: Esquema de simulações dos alinhamentos e otimizações das árvores verdadeiras e alternativas 33
17: Exemplos de padrões em análises de resíduos. 35
18: Verificação gráfica da normalidade 36
19: Dispersão dos valores de $\Delta\ln L$. 44
20: Relação da distância BSD entre árvores alternativas com a média de suas distâncias para a verdadeira correspondente 45
21: Dispersão de ΔlnL com diferentes números de ramos, após redução da amostra. 46
22: Relação entre as variáveis parcial ou totalmente controladas e o desvio padrão (SD) de $\Delta\ln L$ 47
23: Relação entre as variáveis parcial ou totalmente controladas e o desvio padrão (SD) de $\Delta\ln L$ após transformações finais 48
24: Adequação dos resíduos (E) do modelo selecionado aos pressupostos da regressão múltipla 49
25: Proporções GC nas terceiras posições de códons (GC3) dos loci que recuperaram cada uma das três topologias mais frequentes. 52
26: Número de sítios nos loci que recuperaram cada uma das três principais topologias 54
27: ARSs de Afrotheria e Xenarthra nas árvores que obtêm cada uma das três resoluções mais frequentes 55
28: Árvore de máxima verossimilhança para 42 gêneros de Mammalia com base no concatenado de 787 genes nucleares 57
29: Distribuição dos p-valores (via AU) para as árvores T4, T5, T6 e T7, pelos genes que recuperam T1 (Boreoeutheria monofilético e Afrotheria como superordem mais externa) como árvore ML 61
30: Distribuição dos p-valores (via AU) para as árvores T4, T5, T10 e T14, obtidos pelos genes que recuperam T2 (Boreoeutheria monofilético e Xenarthra como superordem mais externa) como árvore ML 62

Lista de quadros e tabelas

Quadro 1: Todas as relações possíveis entre as superordens de Placentalia, em ordem de ocorrência entre as árvores de gene recuperadas, via máxima verossimilhança 51
Tabela 1: Proporções (%) de rejeição da hipótese nula nos testes monocaudais ED e KH. 58
Tabela 2: Proporções (%) de rejeição da hipótese nula nos testes SH e AU 60
Tabela 3: Comparativo entre as proporções de rejeição (%) pelo teste ED e as medianas das distâncias topológicas (BSD) 64

Sumário

Inferindo o desconhecido (prefácio) X

1 – Introdução 1

1.1 – Estimando por Máxima Verossimilhança 1

1.2 – Assumindo erros 2

1.3 – Contemplando incertezas 4

1.4 – Estimando a evolução biológica 6

1.5 – Contemplando incertezas em filogenias: um problema de múltiplas dimensões 8

1.5.1 - O *bootstrap* 10

1.5.2 – Testes topológicos por diferença de verossimilhança 11

1.5.2.1 – KH 12

1.5.2.2 – SH 14

1.5.3 – Correções para o BP como teste topológico 15

1.6 – Erros de estimação na era da Filogenômica 16

1.6.1 – Sinal não-filogenético 16

1.6.2 – Heterogeneidade de sinal entre genes 18

1.7 – O legado dos testes topológicos 20

1.8 – A raiz dos placentários 22

2 – Objetivos 24

2.1 – Geral 24

2.2 – Específicos 24

3 – Metodologia 25

3.1 – Análises da distribuição de $\Delta \ln L$ 25

3.1.1 – Parametrização das simulações e topologias verdadeiras 25

3.1.2 – Obtenção das topologias alternativas 27

3.1.3 – Simulação dos alinhamentos 30

3.1.4 – Otimização de parâmetros e cálculo dos valores de $\Delta \ln L$ 31

3.1.5 – Regressão múltipla dos valores de $\Delta \ln i$ 34

3.2 – Equidistant Delta 37

3.3 – Análises dos dados empíricos 38

3.3.1 – Amostragem das sequências e taxonomia 38

3.3.2 – Processamento das sequências e escolha do modelo de substituição 39

3.3.3 – Inferências filogenéticas e testes topológicos 40

4 – Resultados 43

4.1 – Da influência das variáveis observadas sobre a distribuição de Δln*L* 43

4.1.1 – Análises de dispersão 43

4.1.2 – Análises de regressão e o modelo preditor 46

4.2 – Das resoluções para a raiz de Placentalia pelos genes amostrados 50

4.2.1 – Árvores de genes 50

4.2.2 – Busca por artefatos 51

4.2.3 – Análises concatenadas 56

4.2.4 – Testes topológicos 57

4.3 – Da performance do *Equidistant Delta* 63

5 – Discussão 65

5.1 – Previsibilidade da distribuição de Δln*L* e a viabilidade do ED 65

5.1.1 – Pelo número de sítios 66

5.1.2 – Pela proporção de *gaps* 67

5.1.3 – Pelo número de ramos 68

5.1.4 – Pelas distâncias topológicas 69

5.1.5 – Outras variáveis 70

5.2 – Estimativas pontuais para a raiz de Placentalia 71

5.2.1 – Conteúdo GC, tamanho dos *loci* e acúmulo de substituições 71

5.2.2 – Parametrização das inferências 72

5.3 – Testes topológicos e a significância das estimativas 74

6 – Conclusões 76

Referências 77

Inferindo o desconhecido (prefácio)

Há pelo menos 200 milhões de anos, a curiosidade atenta o *Homo sapiens* aos padrões que permeiam incontáveis fenômenos naturais. Não nos limitou a indagar o 'que acontece?', mas também o desafiador 'como acontece?', ao contemplarmos desde a queda de relâmpagos até as diferentes cores e texturas em ervilhas domésticas. Foi questão de tempo até a matemática se mostrar capaz de dirigir cada uma dessas perguntas através de diferentes abordagens. Hoje, parte dessas abordagens compõe o que conhecemos como estatística descritiva - para o que observamos - e outra parte, a estatística inferencial - para o que vai além do que podemos observar.

A estatística descritiva é a que nos permitiria, em um momento de ócio, resumir *o que acontece* quanto à média e a variação de altura de pessoas alinhadas em uma fila de banco. Por outro lado, a estatística inferencial nos possibilita ir além, por exemplo estimando a altura média da população nacional sem precisar percorrer uma fila de 207 milhões de pessoas; ou mesmo determinando *como acontece* a variação de altura entre humanos através das gerações. Para isso, a estatística inferencial exige tanto a proposição de hipóteses que expliquem o processo incompreendido (ou as medidas populacionais desconhecidos) quanto a avaliação dessas hipóteses dada sua adequação à realidade observada (dados amostrais, conhecidos). À luz dessa lógica, em 1922, Ronald Fisher popularizou o método de estimação por máxima verossimilhança.

Uma pequena parte dos eventos que se sucederam, incluindo todos aqueles que eventualmente motivaram a composição desta dissertação, é narrada na introdução a seguir. Este primeiro capítulo discorre sobre a base teórica necessária para a compreensão do conteúdo nos demais. Se desejar, o leitor familiarizado com os conceitos de verossimilhança, estimação por máxima verossimilhança, tipos de erro estatístico e testes de hipótese pode avançar diretamente para a seção **1.4**. O familiarizado, ainda, com inferência filogenética por máxima verossimilhança, com os diferentes tipos de testes topológicos e suas limitações, pode escolher iniciar da seção **1.6** e, assim por diante, de acordo com cada tema, até a síntese da motivação do trabalho, na seção **1.7**.

1 – Introdução

1.1 – Estimando por Máxima Verossimilhança

Assumindo que um processo qualquer possa ser representado por um modelo probabilístico, os valores (desconhecidos) dos parâmetros nesse modelo podem ser estimados por máxima verossimilhança. Por exemplo, sabendo que o resultado do lançamento de uma moeda pode ser modelado pela função binomial (e.g. sucesso = cara, derrota = coroa), é possível estimar por máxima verossimilhança o valor do parâmetro *p* (probabilidade de obter cara) dessa essa moeda.

A verossimilhança (*L*) de um modelo qualquer variará em função dos valores de seus parâmetros e equivalerá ao produto das probabilidades de se obter cada observação do conjunto em mãos (amostra), dados os valores assumidos para os parâmetros:

$$L(\theta) = \prod_{i=1}^{n} f\,(x_i|\theta)$$

Em que θ é o parâmetro (ou parâmetros) que se deseja estimar, n o número total de observações, x_i a iésima observação realizada e, $f(.)$, a função que descreve a probabilidade da observação.

Dessa forma, os parâmetros de máxima verossimilhança de um modelo probabilístico qualquer são aqueles que maximizam a verossimilhança do modelo diante das observações realizadas (Fisher 1922). No exemplo da moeda, o *p* de máxima verossimilhança será aquele que maximizar a probabilidade - dada pela função binomial - dos resultados (caras e coroas) obtidos para *n* lançamentos da moeda (**Figura 1a**). Como o produto das probabilidades acaba atingindo valores baixíssimos para grandes quantidades de observações, frequentemente usa-se o logaritmo natural das verossimilhanças para facilitar a manipulação dos valores (**Figura 1b**).

Quando formalizado por Fisher em 1922, esse método se contrapôs aos métodos de inferência por "probabilidade inversa" (conhecida hoje como probabilidade posterior, em inferência bayesiana). Ao argumentar a favor da função de verossimilhança como uma base mais apropriada para inferências estatísticas, Fisher contribuiu para a consolidação da estatística frequentista nesse campo, até então dominado pela abordagem bayesiana.

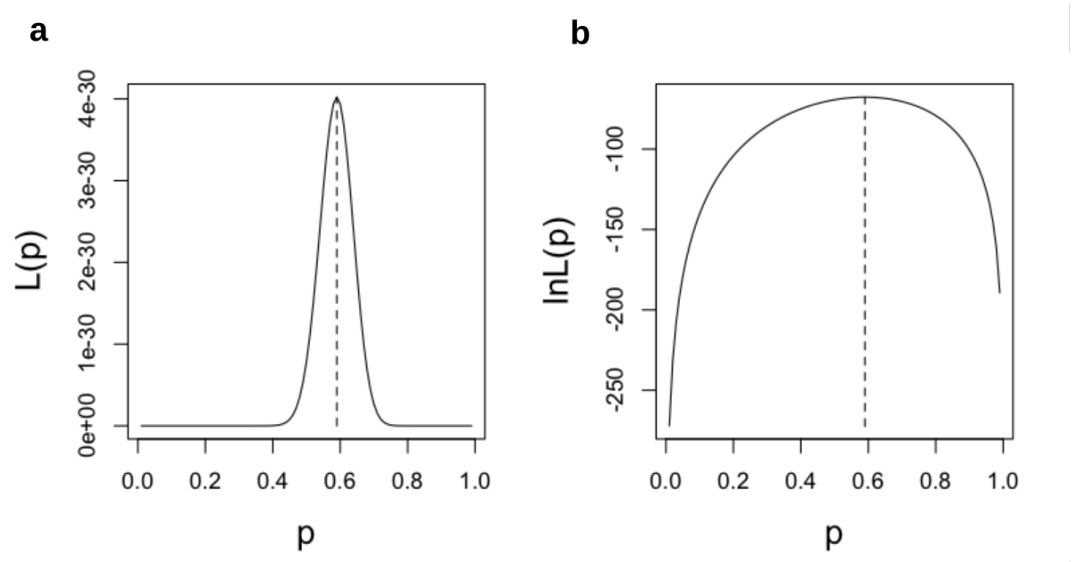


Figura 1: Curva de verossimilhança para a probabilidade de se obter cara (p), dados 100 lançamentos de uma moeda. A estimativa de ML foi p = 0,59, com *L* de aproximadamente 4,02e-30 (**a**) e ln*L* de -67,68 (**b**). Note que a curva de verossimilhança não equivale a uma função de densidade de probabilidade (a área abaixo da curva não totaliza 1), mas a um plot das verossimilhanças para diferentes valores do parâmetro.

1.2 – Assumindo erros

Bem como todo método de inferência estatística, o de máxima verossimilhança está sujeito a erros de estimação causados por falhas ao longo do procedimento experimental. Por isso, em muitos casos é de interesse acessar, além dos parâmetros estimados, a precisão e acurácia das estimativas realizadas. Os erros associados a qualquer inferência podem ser separados em dois tipos de acordo com sua fonte: (1) erro amostral (ou randômico), que é resultado de variabilidade amostral e é gerado com toda e qualquer amostra finita e (2) erro sistemático, que pode ter variadas fontes (exceto variabilidade amostral), normalmente ligadas a desajustes no método de inferência, como por exemplo na presunção de um modelo que não corresponda ao processo que gera os dados observados (Rothman et al. 2008:284).

Como a própria denominação sugere, o erro do tipo randômico diminui a precisão da estimativa, mas não gera viés. Ou seja, somente sob erro randômico, diferentes amostras da mesma população levarão à estimação de diferentes valores para determinado parâmetro, mas esses diferentes valores tenderão a se distribuir em torno do valor real desse parâmetro. Esse tipo de erro tende a desaparecer conforme se aumenta o tamanho da amostra (**Figura 2**).

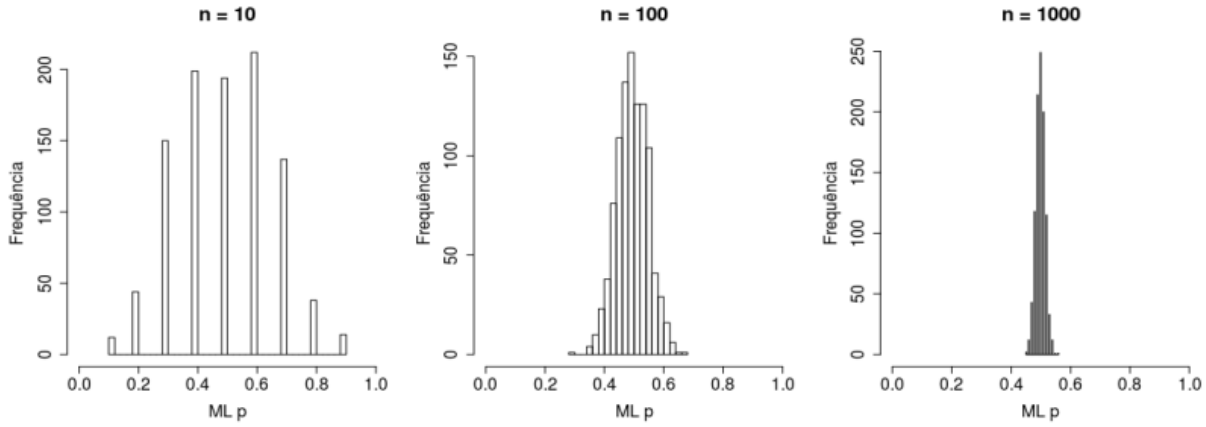


Figura 2*:* **Distribuição amostral de 1000 estimativas independentes de p** por máxima verossimilhança, dados 10, 100 ou 1000 lançamentos da moeda para cada estimativa. À medida que se aumenta o tamanho das amostras, reduz-se o erro amostral, o que reflete numa maior precisão das estimações e menor variância da distribuição.

O erro sistemático, por sua vez, não é aleatório e tende a gerar viés consistentemente. Ou seja, dadas diferentes amostras ele pode, por exemplo, levar à estimação de valores sistematicamente acima do valor real do parâmetro (**Figura 3**). Diferente do erro randômico, o sistemático pode persistir à medida que se aumenta o tamanho da amostra.

Tanto a abordagem bayesiana quanto a frequentista dão acesso à precisão da estimativa; a bayesiana de forma mais direta, a partir de intervalos de credibilidade – nas distribuições de probabilidade posterior de cada parâmetro – e a frequentista, de forma menos trivial, a partir de um intervalo de confiança derivado da matriz de curvatura da função de verossimilhança para o modelo (ver VanderPlas 2014 para diferenças entre intervalo de credibilidade e intervalo de confiança). Portanto, a porção de erro na estimativa de um parâmetro que é do tipo randômico pode ser totalmente acessada pelos próprios recursos do método de inferência. Já a que advém de erro sistemático, em contrapartida, como não necessariamente reduz a precisão da estimativa (apesar de afetar sua acurácia), precisa ser acessada através de abordagens mais sofisticadas.

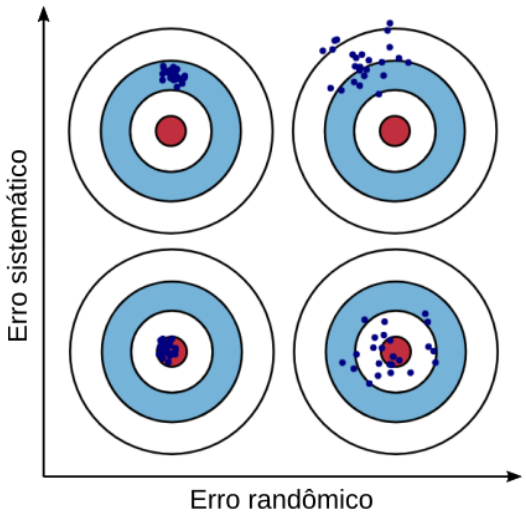


Figura 3: Precisão e acurácia das estimativas na presença de erro amostral vs erro sistemático. Assuma-se que os pontos sejam as estimativas para um parâmetro dadas N amostras e o centro do alvo sejo o valor verdadeiro do parâmetro; o erro amostral (ou randômico) reduz a precisão das estimativas e o erro sistemático gera viés, reduzindo sua acurácia. A prevalência de cada efeito depende da quantidade gerada de cada tipo de erro. Figura adaptada de scott.fortmann-roe.com/.

1.3 – Contemplando incertezas

Por estar potencialmente ligado a diferentes causas e não ser facilmente identificável pelo próprio método de estimação, o erro sistemático é um dos maiores desafios enfrentados por estudos que passam por qualquer inferência estatística. Particularmente, o erro que se gera ao assumir modelos probabilísticos inadequados sempre foi de interesse em estatística inferencial; na qual, ao longo do século XX, desenvolveram-se múltiplas formas de testar a adequação de modelos discrepantes a partir de seus valores de verossimilhança.

Neyman e Pearson (1933) iniciaram esse legado ao empregarem, pela primeira vez, a razão de verossimilhança (LR – *likelihood ratio*) entre modelos para comparar sua adequação aos dados. Pouco depois, Wilks (1938) demonstrou que, quando um par de modelos é hierárquico (é possível obter um restringindo um ou mais parâmetros no outro), a razão entre suas verossimilhanças tem distribuição amostral previsível. Ele provou que, conforme o tamanho das amostras se aproxima de infinito (condição assintótica), o logaritmo natural desses LRs, multiplicado por menos dois (-2lnLR), tende a se distribuir de acordo com uma função de χ^2, com número de graus de liberdade igual à diferença entre o número de parâmetros livres em cada modelo.

A conclusão de Wilks (1938) desencadeou o advento e amplo uso de um tipo de teste que ficou conhecido como LRT (*likelihood ratio test*), principalmente para averiguar se a adição de mais parâmetros aos modelos assumidos aumentava significativamente sua

adequação aos dados. Nesse teste, calcula-se, por exemplo, a máxima verossimilhança L*f* de um modelo com função *f*(P1,P2) e a máxima verossimilhança L*g* de um modelo mais complexo, *g*(P1,P2,P3,P4). Se o valor resultante de -2*log(L*f*/L*g*) ultrapassa o limite de confiança escolhido (e.g. 95%) na distribuição de χ^2 com dois graus de liberdade (4 parâmetros em *g* menos 2 parâmetros em *f*), o modelo *g* mostra-se significativamente mais verossímil, portanto rejeita-se o modelo mais simples (**Figura 4**).

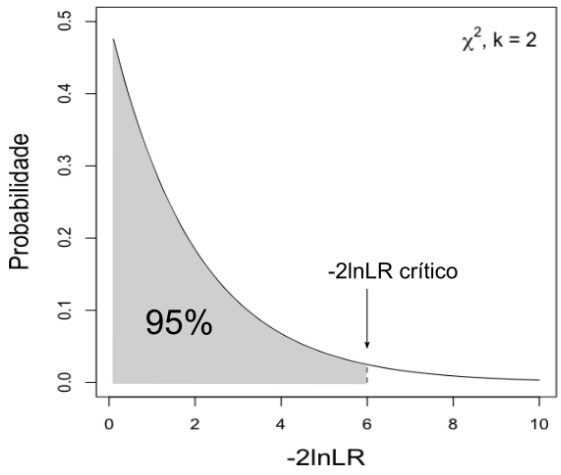


Figura 4: Intervalo de confiança do LRT dado pela função de χ^2 com 2 graus de liberdade. Se -2lnLR entre o modelo f e o modelo g ultrapassar a região cinza indicada, rejeita-se o modelo f com 95% de confiança. Do contrário, não pode-se afirmar que o modelo g seja significativamente mais adequado.

Ao estender a abordagem de Wilks para outras famílias de modelos probabilísticos, Cox (1961) demonstrou que, em condições assintóticas, os LRs entre dois modelos não-hierárquicos, por sua vez, teriam distribuição normal. Entretanto, por fixarem um dos modelos como a hipótese nula do teste, os procedimentos de Wilks e Cox assumem, implicitamente, que o verdadeiro está entre os candidatos. Nos casos em que isso não procede, o LRT pode ser inválido (White, 1982).

Alternativamente, após quase três décadas, Vuong (1989) veio a propor testes de hipótese baseados na LR entre pares de modelos probabilísticos equidistantes do modelo verdadeiro; uma abordagem não só robusta, mas *voltada* para casos em que há especificação de modelos inadequados. Ao derivar matematicamente a distribuição assintótica do LR entre modelos lineares em diversas condições, ele demonstra que, na condição de equidistância e entre modelos hierárquicos, a distribuição assintótica de LR se mantém conforme Wilks (1938). Já entre modelos não-hierárquicos ou sobrepostos, aproxima-se de uma distribuição

normal, mas sendo necessárias correções quando o número de parâmetros difere entre modelos.

1.4 – Estimando a evolução biológica

A estatística inferencial teve notável utilidade, conquistando destaque histórico muito cedo entre as ciências biológicas, principalmente na área da epidemiologia, em que há registros de estimações estatísticas desde 1662, com o trabalho de John Graunt no livro *Natural and Political Observations Made upon the Bills of Mortality*. Mais recentemente, os trabalhos de Ronald Fisher (1918, 1930), Sewall Wright (1932), Theodosius Dobzhansky (1937) e Ernst Mayr (1942), entre outros, foram essenciais ao fornecer a base teórica necessária para os estudos em genética de populações e evolução biológica; contribuição que resultou na síntese evolutiva moderna e sua eventual formalização estatística.

Os principais e mais utilizados modelos probabilísticos de evolução molecular, entretanto, só surgiram na segunda metade do século XX. Baseados em cadeias de Markov de tempo contínuo, esses modelos descrevem matrizes de transição (lê-se mudança) com probabilidades de troca de uma base (A,C,T ou G) por outra, em função do tempo (**Figura 5**). Um dos primeiros e mais simples modelos do tipo foi o proposto por Jukes e Cantor (1969) assumindo que essas probabilidades seriam iguais entre qualquer base e que as proporções das quatro bases em uma sequência (frequências de equilíbrio) tenderiam a se igualar (25% cada) ao longo do tempo. Dependendo dos pressupostos de cada modelo de substituição, permitindo que mais ou menos desses parâmetros variem livremente, assume-se uma matriz de taxas de substituição (Q) e um vetor de frequências de equilíbrio diferentes, para obter a matriz de transição (P) esperada.

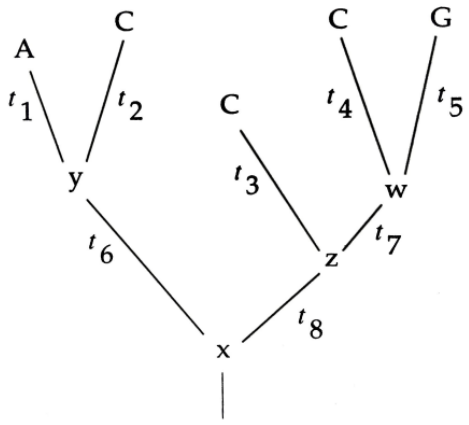
$$P(t) = \begin{pmatrix} p_{AA}(t) & p_{GA}(t) & p_{CA}(t) & p_{TA}(t) \\ p_{AG}(t) & p_{GG}(t) & p_{CG}(t) & p_{TG}(t) \\ p_{AC}(t) & p_{GC}(t) & p_{CC}(t) & p_{TC}(t) \\ p_{AT}(t) & p_{GT}(t) & p_{CT}(t) & p_{TT}(t) \end{pmatrix}$$

Figura 5: Matriz de probabilidades de transição (mudança) entre bases em função do tempo. Por exemplo, $p_{AT}(t)$ é a probabilidade de mudança de C para T ao longo de t e $p_{AA}(t)$ é a probabilidade de que A permaneça o mesmo. Essas probabilidades variarão de acordo com os parâmetros do modelo de substituição assumido. Matriz retirada de en.wikimedia.org/.

Assumindo que o processo evolutivo em nível molecular seja reversível (e.g. a probabilidade de mudança de A para G seja igual a de G para A), os modelos de substituição baseados em processos de Markov também podem ser restritos para refletir essa reversibilidade. Assumir essa propriedade nos permite inferir, por exemplo, a distância evolutiva entre duas sequências (i.e. o número de substituições entre elas) a partir das diferenças entre as bases de cada uma. Pelo mesmo princípio é possível inferir a sequência ancestral de duas sequências existentes ou a ordem de diversificação entre múltiplas sequências (filogenia). Entretanto, o desafio de trazer essa última abstração à prática só foi superado mais recentemente.

Os primeiros modelos de evolução molecular não foram imediatamente implementados para estimativa de árvores evolutivas, tanto pela difícil computação do problema - ainda nos primórdios da era da informática - quanto pela limitada disponibilidade de dados empíricos na forma de sequências de DNA (Felsentein 1981). Da inferência por máxima parcimônia de Edwards & Cavalli-Sforza (1964), passando pela estimação via similaridade par-a-par entre sequências (Fitch & Margoliash 1967), e pelos métodos de compatibilidade (Le Quesne 1969), perdurou uma carência de abordagens probabilísticas para recuperar árvores filogenéticas. Apenas em 1981 Felsenstein desenvolveu uma estratégia que implementa modelos de Markov para computar a verossimilhança de árvores filogenéticas de forma eficiente e utilizando-se diretamente da informação contida em sequências (**Figura 6**).

A partir de então, tornou-se possível inferir uma árvore de máxima verossimilhança ao encontrar os parâmetros do modelo de substituição, a topologia (estrutura hierárquica da árvore) e os comprimentos dos ramos na árvore que maximizam sua função de verossimilhança, com o alinhamento de sequências em mãos. Por permitir a acomodação dos comprimentos de ramo ao longo de cada árvore possível até atingir máxima verossimilhança, o método de Felsenstein permitiu maior precisão na estimação da história de grupos com diferentes taxas de mudanças entre suas linhagens - fenômeno antes problemático para os métodos de parcimônia e compatibilidade (Felsenstein 1978). Com isso, o método de máxima verossimilhança foi rapidamente popularizado em sistemática molecular.



$$L = \text{Prob}(D|T) = \prod^{m} \text{Prob}\left(D^{(i)}|T\right)$$

$$\text{Prob}\left(D^{(i)}|T\right) = \sum_x \sum_y \sum_z \sum_w \text{Prob}(A, C, C, C, G, x, y, z, w|T)$$

Figura 6: Árvore filogenética e função de verossimilhança correspondente. A verossimilhança de uma árvore filogenética (T) é calculada independentemente para cada sítio ($D^{(i)}$), sendo resultado do produto entre as probabilidades de cada ramo na árvore. Por sua vez, a probabilidade do ramo é dada pela matriz de transição, sendo a soma das probabilidades de transição de cada estado possível (A,C,T ou G) no nó ancestral (x,y,z ou w) para o nó ou terminal subsequente. Por fim, a verossimilhança da árvore dado o alinhamento é o produto dos valores de verossimilhança obtidos para cada um dos *m* sítios (ou somatório, se utilizadas log-verossimilhanças). Árvore e fórmulas retiradas do livro *Inferring Phylogenies* (Felsenstein 2004).

Apesar de consistente (quando assume-se o modelo evolutivo correto, isto é, topologia e modelo de substituição apropriados) e relativamente robusto à violação do modelo de substituição (Huelsenbeck 1995), o método ainda está sujeito a erro sistemático e a erro amostral, que, por sua vez, se agrava quanto maior a complexidade dos modelos assumidos (Yang 2006:37). Além disso, o número de topologias possíveis cresce fatorialmente com o número de ramos (Felsenstein 1978), frequentemente sendo necessários métodos de busca heurística (via rearranjo topológico) para encontrar a árvore de máxima verossimilhança (ML) sem precisar calcular a de todas as árvores possíveis. Isto torna a inferência da árvore ML ainda mais suscetível a erro sistemático, como por aprisionamento em máximos locais (Chor et al. 2000). Assim, surge uma demanda por meios de testar a adequação de modelos evolutivos discrepantes.

1.5 – Contemplando incertezas em filogenias: um problema de múltiplas dimensões

A adequação de modelos de substituição pode ser avaliada pelo LRT de Wilks, testando a melhora no modelo mediante adição de mais parâmetros livres (diferentes taxas de

substituição ou frequências de bases) (Huelsenbeck & Crandall 1997). Apesar dessa abordagem ter sido rapidamente popularizada e implementada em programas altamente citados até hoje, como jModelTest (Posada 2008), já não seria apropriada para teste de topologias com o mesmo número de ramos, por exemplo. Uma vez que topologias não podem ser representadas por um único parâmetro - apenas pelos múltiplos parâmetros que são os comprimentos de seus ramos - duas topologias com mesmo número de ramos não teriam grau de liberdade entre elas, impossibilitando o teste via χ^2.

Ademais, qualquer tentativa de delimitar um intervalo de confiança (IC) para topologias teria de lidar com a complexidade de um espaço paramétrico com uma dimensão para cada ramo possível, e diferentes topologias como regiões discretas nesse espaço (**Figura 7a**). Ao longo do espaço, podem haver múltiplos picos de verossimilhança, em regiões (topologias) distintas (**Figura 7b**).

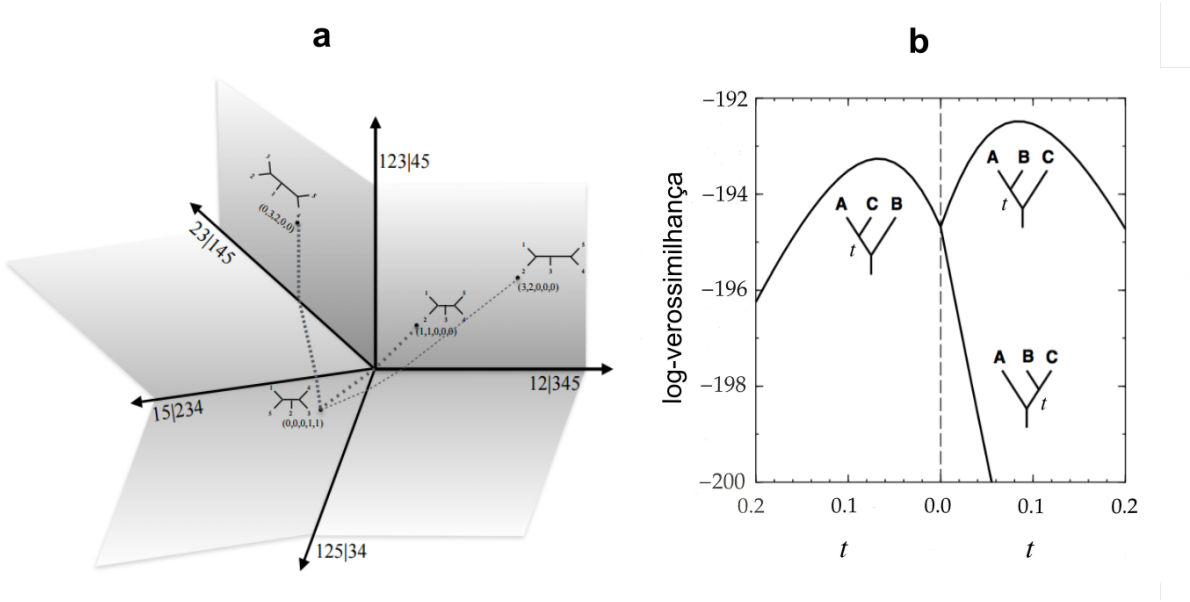


Figura 7: Variação da topologia e da verossimilhança da árvore conforme mudam seus ramos internos. (a) é uma abstração do espaço topológico para cinco táxons. Cada plano corresponde a uma topologia específica e é definido pelos eixos de variação de comprimento dos dois ramos internos na árvore. Por sua vez, (b) coloca a posição e comprimento de um ramo interno t em perspectiva para exemplificar como diferentes resoluções podem gerar picos de verossimilhança diversos ao longo do espaço topológico. Espaço topológico retirado de St. John (2016) e curvas de verossimilhança adaptadas do livro *Inferring Phylogenies* (Felsenstein 2004).

Assim, pode-se derivar um IC a partir da própria função de verossimilhança para a árvore de máxima verossimilhança (ML) encontrada, por exemplo; mas esse intervalo abarcaria os valores aceitáveis de parâmetros (comprimentos de ramos e taxas de substituição) apenas para a topologia em questão – no caso, a da árvore ML. Assim, é possível acessar o efeito do erro randômico na estimação dos parâmetros para a árvore ML pela própria função

de verossimilhança da árvore, mas não comparar sua adequação à de topologias distintas (Felsenstein, 2004:319). Enquanto a abordagem bayeasiana oferece uma solução natural para esse problema, pelo intervalo HPD (*highest probability densisty*) da sua distribuição posterior para a topologia; o desafio de construir um IC a partir da informação obtida no processo de busca heurística pela árvore ML permanece em aberto (Yang 2006:178), sendo necessárias estratégias alternativas para comparar estatisticamente a topologia estimada a hipóteses alternativas.

1.5.1 - O *bootstrap*

A partir dos anos 80, variadas estratégias foram propostas para acessar o erro na especificação de topologias. Uma das primeiras e talvez a mais conhecida tentativa de atribuir confiança estatística aos clados (grupos de táxons com um ancestral em comum e exclusivo, ou monofiléticos) de uma árvore estimada foi o suporte de *bootstrap*, proposto pelo próprio Felsenstein, em 1985.

Baseada no *bootstrap* não-paramétrico de Efron (1979) para obter a distribuição amostral de estimativas de um parâmetro, a versão de Felsenstein (1985) para filogenias é computada a partir de reamostras aleatórias (e com reposição) dos sítios (colunas) do alinhamento de sequências original. Após obter-se, tipicamente, centenas de reamostras de alinhamentos réplicas, cada uma do mesmo tamanho do alinhamento original, são utilizadas para estimar novas árvores réplicas (pseudo-réplicas de *bootstrap*) (**Figura 8**). Finalmente, a proporção (ou suporte) de *bootstrap* para cada clado na árvore estimada originalmente é a proporção de vezes em que o clado aparece entre as pseudo-réplicas.

Apesar de Felsenstein ter inicialmente sugerido a proporção de *bootstrap* como uma medida de **repetibilidade**, ou seja, da tendência de diferentes amostras geradas pelo mesmo processo evolutivo que o alinhamento original a darem suporte aos mesmos clados; muitas vezes ele é interpretado como uma medida da **acurácia** da estimativa, ou da probabilidade do clado estar presente na árvore verdadeira. Essa última interpretação poderia conduzir um pesquisador a medir a acurácia de toda uma topologia baseando-se na proporção de vezes que esta aparece entre as pseudo-réplicas de *bootstrap* (BP – *bootstrap proportion*). Em outras palavras, este viria a utilizar a BP da topologia estimada como um p-valor na hipótese de que ela corresponde à topologia verdadeira.

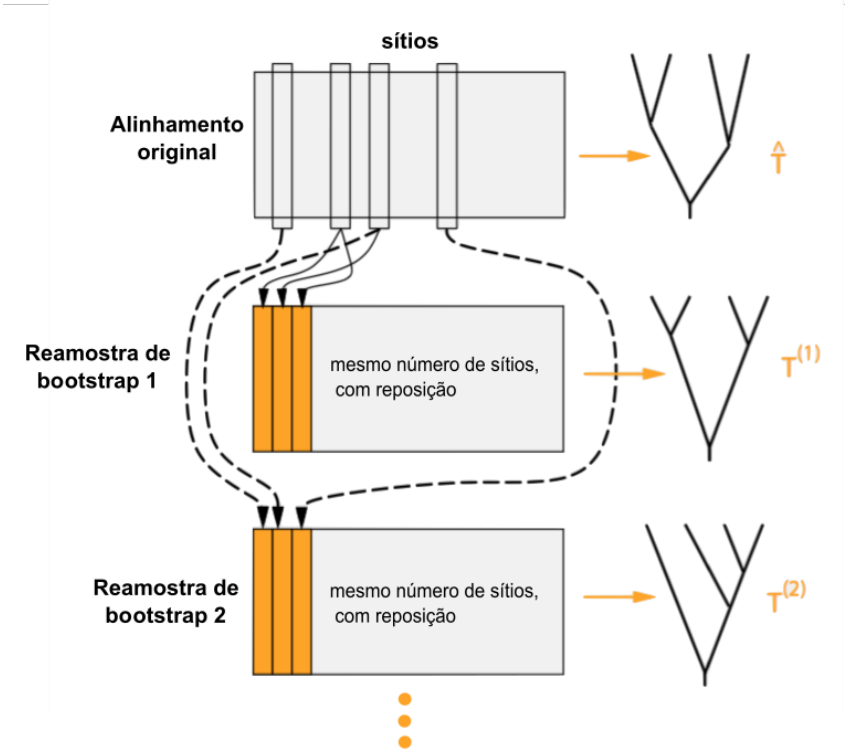


Figura 8: Bootstrap de Felsenstein (1985). A partir do alinhamento original, obtém-se cada reamostra de bootstrap sorteando o mesmo número de sítios aleatoriamente e com reposição. Ou seja, em cada reamostra, pode haver duas ou mais cópias de um sítio específico, bem como nenhuma de outro. Isso garante a variação necessária para acessar a repetibilidade da estimativa original (T^), que, por sua vez, é verificada estimando uma pseudo-réplica ($T^{(i)}$) de cada reamostra. Esquema adaptado de phylo.bio.ku.edu/.

Entretanto, Hillis e Bull (1993) observaram que a medida de acurácia por BP aparenta ser progressivamente enviesada, subestimando cada vez mais o p-valor da topologia estimada quanto mais acurada ela é. Em resposta, Felsenstein e Kishino (1993) demonstraram que o *bootstrap* na verdade não é sistematicamente enviesado para valores mais baixos (sempre conservador), já que a distorção para valores abaixo de 50 tende a se inverter, passando o p-valor a ser progressivamente superestimado. Argumentam, assim, que o problema não está associado à técnica, mas à interpretação do BP diretamente como uma medida da probabilidade (ou p-valor) do clado ou topologia ter sido corretamente estimado.

1.5.2 – Testes topológicos por diferença de verossimilhança

No período que vai do final dos anos 80 ao início dos 2000 houve uma profusão de novas ideias para medir a confiabilidade de filogenias, buscando preencher o *gap* existente. Entre elas, destaco os testes baseadas no delta de verossimilhança, que são amplamente empregados após inferência de filogenias por máxima verossimilhança e são o foco desta dissertação. O valor de 'diferença entre log-verossimilhanças' ($\Delta \ln L$ ou 'delta de verossimilhança') entre duas árvores filogenéticas é equivalente ao log da razão entre suas

verossimilhanças (log(LR) ou ln*L*R), utilizado no LRT. Aqui, passo a referi-los por Δln*L*s para diferenciar os testes que serão apresentados deste outro.

Analogamente ao LRT, os testes de Δln*L* visam determinar se a diferença na adequação entre topologias discrepantes está dentro do esperado por efeito de erro randômico ou se é produto de uma diferença significativa de adequação (erro sistemático) entre elas. Todos eles se diferenciam dos LRTs por não assumirem de antemão a variância esperada para o Δln*L* (dada pelo número de graus de liberdade no LRT). Alternativamente, para acessar essa variância e determinar um intervalo de confiança para Δln*L* sob sua hipótese nula, esses testes se utilizam de variadas abordagens; das que fazem pouca ou nenhuma suposição *a priori* em relação à distribuição de Δln*L*, sob custo de maior esforço computacional, até as que dependem de mais pressupostos, mas simplificam ao máximo a realização do teste.

1.5.2.1 – KH

O teste proposto por Kishino & Hasegawa (1989) foi o primeiro a utilizar o Δln*L* entre filogenias. Na forma como foi originalmente engendrado e implementado no pacote de programas PHYLIP (Felsenstein 1989), o teste KH invoca o teorema do limite central para assumir que (1) a distribuição amostral de Δln*L* entre filogenias é assintoticamente normal e (2) a variância dessa distribuição pode ser derivada da variância dos valores Δln*L* obtidos para cada sítio do alinhamento (Kishino & Hasegawa 1989). Assim, multiplicando a raiz da variância (desvio padrão) pelos quantis de interesse na normal padrão, obtém-se o intervalo de confiança desejado (**Figura 9**). Se o valor original de Δln*L* entre as topologias em teste cair fora desse intervalo, rejeita-se a hipótese nula de que ambas sejam explicações igualmente razoáveis da evolução do alinhamento.

Entretanto, não há comprovação de que ambos os pressupostos 1 e 2 do KH original se sustentem em qualquer caso. Como uma alternativa mais custosa computacionalmente, os próprios autores do teste KH haviam sugerido que a distribuição de Δln*L* poderia ser obtida por um procedimento de reamostragem baseado no *bootstrap* (Hasegawa & Kishino 1989). Nesse caso, em vez de utilizados para estimar novas árvores, as reamostras de *bootstrap* serviriam para otimizar (reestimar), via ML, todos os parâmetros (de substituição e comprimentos de ramos) para ambas as árvores em teste, fixando-se suas topologias. Em seguida, os valores de verossimilhança obtidos de cada reamostra, para cada topologia, comporiam a distribuição amostral de Δln*L*, da qual poderia se obter um intervalo de confiança diretamente.

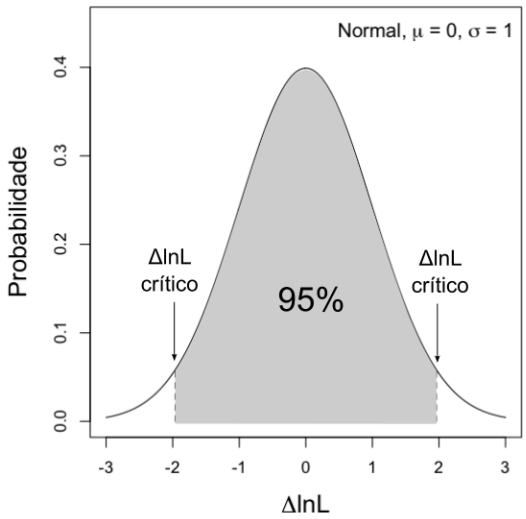


Figura 9: Intervalo de confiança (IC) do teste KH pressupondo a normalidade da distribuição de Δln*L*. Para saber se a diferença de verossimilhança entre duas árvores é significativa com 95% de confiança, deriva-se a variância amostral a partir da variância dos Δln*L* por sítio e multiplica-se a raiz dessa variância (desvio padrão) pelos quantis 2,5% e 97,5% da normal padrão (-1,96 e 1,96, respectivamente) para obter os limites inferior e superior do IC.

Sendo bem mais exigente computacionalmente, a abordagem via *bootstrap* vem a custo de maior tempo de processamento. Buscando outra alternativa para o problema, fora a presunção de normalidade, Kishino e colaboradores (1990) demonstram que uma reamostragem poderia ser realizada diretamente dos valores de verossimilhança por sítio estimados para cada árvore, no lugar dos próprios sítios do alinhamento. Essa técnica, que ficou conhecida como RELL (*resampling estimated log-likelihoods*), acelera a realização do teste ao dispensar a otimização dos mesmos parâmetros a cada reamostragem, mas também depende de condições assintóticas (grandes amostras e modelo de substituição corretamente especificado) (Goldman et al. 2000).

Ainda, qualquer aplicação do KH promove direta ou indiretamente uma centralização dos valores de Δln*L*, assumindo que a média de sua distribuição seja 0, conforme a hipótese nula. Esse não é um pressuposto razoável, dado que o próprio teste topológico é motivado pela desconfiança de que uma das topologias seja mais adequada aos dados que a outra. Desejando-se testar, por exemplo, se a topologia de ML é significativamente melhor que qualquer outra, não haveria por que supor que a distribuição amostral de seus Δln*L* tenha média 0 ou que a variância obtida para a distribuição (assumindo normalidade ou não) permaneceria a mesma sob a hipótese nula. Essa distorção, chamada viés de seleção, invalida os resultados do KH (Shimodaira & Haseagawa 1999; Goldman et al. 2000).

1.5.2.2 – SH

Buscando solucionar o viés de seleção dos casos em que a árvore ML, identificada *a posteriori*, é inclusa no teste, Shimodaira e Haseagawa (1999) desenvolveram uma modificação do KH para permitir múltiplas comparações entre árvores. Conhecida como teste SH, essa modificação compara simultaneamente todas as árvores em um conjunto dado *a priori* sob a hipótese nula de que todas explicam igualmente bem a evolução do alinhamento.

De forma análoga ao KH, o teste se inicia obtendo os valores de $\Delta\ln L$ entre as árvores testadas, mas, no SH, são avaliados apenas os $\Delta\ln L$s entre a árvore de maior verossimilhança no conjunto (ML) e cada uma das demais. Em seguida, como no KH via *bootstrap*, podem ser realizadas múltiplas reamostragens do alinhamento original e, para cada uma, otimizados os parâmetros de substituição e comprimentos de ramo de cada árvore. Em sua publicação original, entretanto, Shimodaira e Hasegawa (1999) implementam a técnica RELL no lugar do *bootstrap* e otimização completa. Por um meio ou por outro, esse procedimento dá origem a uma matriz de valores de log-verossimilhança ($\ln L$), em que cada fileira corresponde a uma das árvores no conjunto e cada coluna a uma reamostra de *bootstrap*.

No SH, o procedimento de centralização também é realizado; todos os valores de $\ln L$ na matriz são subtraídos pela média de sua fileira, para garantir que os dados reamostrados se adequem à hipótese nula. Então, em cada coluna da matriz ajustada é selecionado o maior valor de $\ln L$, ou seja, a verossimilhança da árvore ML para a respectiva reamostra (que não necessariamente corresponderá à ML para o alinhamento original). Com isso, forma-se uma nova matriz, composta pelos valores $\Delta\ln L$ entre os os $\ln L$ "vencedores" de cada coluna e os demais (é assim que o teste permite a seleção *a posteriori* da árvore ML). Finalmente, as fileiras dessa nova matriz corresponderão à distribuição de $\Delta\ln L$ esperada entre a árvore ML original e cada árvore do grupo, no caso de serem igualmente adequadas para explicar a evolução do alinhamento.

Note que, diferente do KH, este trata-se de um teste monocaudal, uma vez que qualquer valor de $\Delta\ln L$ entre a ML para cada amostra e outra árvore será necessariamente maior do que 0. Assim, se um ou mais dos valores de $\Delta\ln L$ computados inicialmente caírem fora do IC descrito pela distribuição monocaudal correspondente, rejeita-se a hipótese nula de que todas as árvores são igualmente adequadas. Além disso, pode-se determinar um p-valor para cada $\Delta\ln L$ pela posição em sua respectiva distribuição e, assim, delimitar um IC composto por aquelas árvores que obtiveram p-valor acima do limite de significância escolhido, sendo igualmente adequadas para o SH.

Entretanto, para garantir que os níveis de significância computados nesse tipo de teste sejam acurados, é preciso que o grupo de topologias testadas compreenda todas as que possam ser eventualmente aventadas como a ML para dada amostra de *bootstrap* (Westfall & Young 1993:48). Shimodaira & Hasegawa (1999) argumentam que não é necessário incluir todas as árvores possíveis, sendo desejável manter as mais improváveis fora do conjunto. Na prática, o teste SH mostra-se bastante conservador se comparado aos demais testes de $\Delta \ln L$, incluindo cada vez mais árvores em seu IC quanto maior é o grupo de árvores testadas e exigindo alinhamentos de tamanho cada vez maior para equiparar seu poder de rejeição ao de outros testes (Strimmer & Rambaut 2001).

1.5.3 – Correções para o BP como teste topológico

Seguindo as discussões sobre a validade da proporção de *bootstrap* como uma medida de acurácia topológica, Efron e colaboradores (1996) demonstraram que a distorção entre a BP e o p-valor correto para a árvore, observada anteriomente por Hillis e Bull (1993), é produto de curvaturas irregulares nas fronteiras entre as diferentes topologias no espaço paramétrico. Como solução para a distorção, propuseram uma aproximação mais adequada ao p-valor através de um algoritmo que aplica tanto *bootstrap* não-paramétrico (visto em **1.5.1**) quanto paramétrico (replicação do alinhamento por simulações), em duas etapas (Efron et al. 1996). No entanto, por exigir a geração de um grande número de réplicas para uma aproximação adequada, esse método representa um fardo computacional muito maior do que o do BP convencional.

Em 2002, partindo da mesma abordagem teórica de Efron, Shimodaira desenvolveu um teste topológico que transforma o BP em um p-valor mais adequado ao estimar a curvatura do espaço topológico nas regiões de fronteira das múltiplas topologias comparadas e suas distâncias em relação a cada árvore. Apesar do teste SH também ser apropriado para múltiplas comparações, o teste de Shimodaira (2002) mostra-se menos conservador que o SH, mantendo ao mesmo tempo baixas taxas de erro do tipo I (rejeição da hipótese nula - de igual adequação - quando a árvore é menos adequada que as demais) e do tipo II (não rejeição da hipótese nula quando não é verdadeira); motivo pelo qual é nomeado *approximately unbiased* (AU). Entretanto, em vez do *bootstrap* em dois níveis de Efron e coautores (1996), o AU lança mão de reamostragens de *bootstrap* não-paramétrico com variados tamanhos em relação ao do alinhamento original (o *multiscale bootstrap*). Por isso, apesar de sua maior acurácia, o AU exige o maior esforço computacional entre os testes apresentados aqui.

1.6 – Erros de estimação na era da Filogenômica

À medida que os métodos de sequenciamento se tornaram progressivamente eficientes e o preço por base de DNA sequenciada passou a cair rapidamente (ver Dijk et al. 2014), a quantidade e diversidade taxonômica das sequências disponíveis para análises filogenéticas também aumentou, dando início à era da Filogenômica. Contando com conjuntos de dados cada vez maiores, os trabalhos nessa nova área vem se tornando regra, alimentados pela expectativa de se resolver, enfim, persistentes problemas macroevolutivos. Entretanto, tais expectativas vêm perdendo força conforme diferentes conjuntos de dados, formas de processamento e métodos de inferência recuperam resoluções distintas para alguns desses problemas (e.g. filogenia de eumetazoários).

Um pesquisador pode vir a esperar respostas definitivas da análise de múltiplos marcadores concatenados por saber que mais observações tendem a reduzir o erro amostral associado a estimações estatísticas (**Figura 2**). Entretanto, além do erro amostral não ser o único erro possível, como visto anteriormente, essa esperança é ainda menos realista em inferência filogenética por dois motivos: (1) a adição de mais genes pode levar não apenas ao aumento de sinal filogenético (similaridades resultantes da história compartilhada pelas espécies) como também de sinal não-filogenético (ruído) e (2) diferentes genes podem ter histórias evolutivas discrepantes, adicionando sinal conflitante ao conjunto de dados.

1.6.1 – Sinal não-filogenético

Definido por Phillipe e colaboradores (2011) como a combinação de diferentes tipos de ruído que competem com o sinal filogenético no processo de estimação de árvores evolutivas, o sinal não-filogenético ocorre quando erros na seleção ou análise dos dados são interpretados como evidências sobre o processo de evolução do grupo de interesse. Caso se acumule no conjunto de dados sem ser propriamente identificado, o sinal não-filogenético tem a perigosa propriedade de levar à inferência de árvores errôneas, mas com alto suporte estatístico (Jeffroy et al. 2006).

Das suas três principais fontes, a primeira se dá na etapa da identificação de sequências ortólogas e a segunda no alinhamento dessas sequências. Em ambos, tenta-se estabelecer relações de homologia por especiação entre as sequências amostradas (ortologia), mas falhas de interpretação em uma ou outra etapa podem tomar por ortologias o que são, na verdade, paralogias (homologias por duplicação) entre sequências (**Figura 10**) ou homoplasias (similaridade por convergência) entre bases.

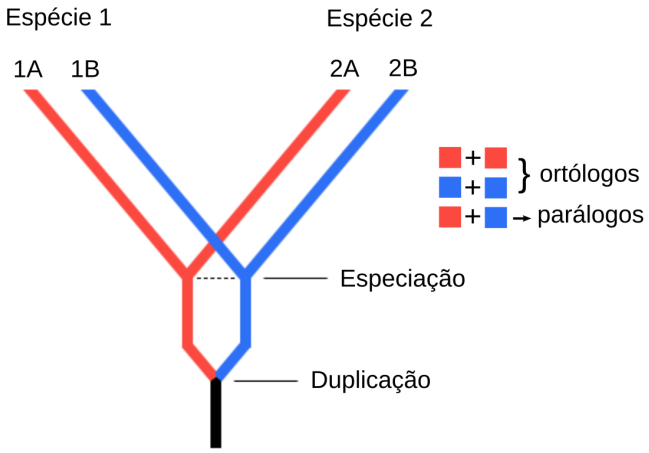


Figura 10: Cópias ortólogas e parálogas ao longo da história evolutiva. Sequências ortólogas têm sua história em comum interrompida no evento de especiação das linhagens de organismos que as contém; em contrapartida, sequências parálogas de espécies distintas já tinham histórias independentes no momento da especiação, tendo MRCA antes de um evento de duplicação gênica anterior. Figura adaptada de ecoevo.unit.oist.jp/.

Por sua vez, a terceira das principais fontes de sinal não-filogenético é um erro de natureza sistemática, em que o método de inferência ou modelo de substituição assumido falha em inferir múltiplos eventos de substituição em um mesmo sítio ao longo do tempo. As chamadas sequências saturadas, com muitos desses sítios de evolução rápida, tendem a acumular algumas similaridades ao acaso (homoplasias) entre si, causando o conhecido artefato de atração de ramos longos na filogenia inferida (Bersten 2005).

Apesar de um maior refinamento das técnicas de alinhamento de múltiplas sequências potencialmente evitar a inferência de homologia - alinhamento - entre bases homoplásticas, não são capazes de diferenciar sequências parálogas (homólogas por duplicação) de sequências ortólogas (homólogas por especiação), necessitando a utilização de técnicas mais sofisticadas para seleção de marcadores verdadeiramente ortólogos (ver Vallender 2009). Ainda assim, mesmo que os marcadores analisados sejam de fato ortólogos e suas sequências sejam alinhadas corretamente, eventualmente haverá, por exemplo, taxas heterogêneas de substituição entre linhagens, levando à subestimação ou superestimação do comprimento de alguns ramos, caso o modelo assumido seja simplista demais para contemplar essa característica.

Nesses casos, medidas cabíveis passam por incluir mais espécies no alinhamento, para ‘quebrar’ os ramos de comprimentos mais longos - mais frequentemente subestimados - na árvore ou implementar modelos de substituição mais complexos. Neste último, assumem-se diferentes taxas de substituição ao longo do tempo (heterotaquia) (Kolaczkowski & Thornton

2008), diferentes matrizes de substituição entre os sítios do alinhamento (Lartillot & Philippe 2004) ou, também, combinam-se múltiplas estratégias, acomodando a heterogeneidade evolutiva ao longo do alinhamento e da árvore (Blanquart & Lartillot 2008).

Outra etapa essencial quando considera-se utilizar modelos mais complexos é a realização dos testes de hipótese necessários para verificar sua adequação, já que quanto mais parâmetros inclui o modelo, maior tende a ser o acúmulo de erro amostral (Yang 2006:37). Alinhamentos mais longos tendem a contrabalancear esse acúmulo, comportando modelos mais paramétricos; em contrapartida, quando incluem múltiplos genes, pode ocorrer conflito de sinais.

1.6.2 – Heterogeneidade de sinal entre genes

Ocasionalmente, árvores recuperadas a partir de diferentes genes (árvores de genes) podem diferir entre si e da árvore que verdadeiramente representa os padrões de especiação na história do grupo em estudo (árvore de espécies). Porém, nem sempre isso ocorre por erros de estimação; em alguns casos, o verdadeiro processo de evolução do gene de fato discorda do processo de evolução das espécies das quais foi amostrado (Nichols 2001). Apesar desse fenômeno já ter sido há bastante tempo associado a duplicações gênicas não detectadas - quando genes parálogos são tomados por ortólogos - (Fitch 1970), mais recentemente Madison (1997) discutiu outras fontes possíveis para a incongruência entre árvores de genes, como a transferência horizontal de genes (HGT – *horizontal gene transfer*) e a amostragem incompleta de linhagens (ILS – *incomplete lineage sorting*).

Fenômenos como o HGT reintroduzem cópias de genes a linhagens das quais já foram separados anteriormente por um evento de especiação. Por gerarem reticulações na árvore de espécies, causam distorções na topologia estimada por métodos de inferência filogenética existentes, que contemplam apenas a transferência vertical de genes (via herança). A introgressão (incorporação de cópias genéticas oriundas de hibridização), mais comum em eucariotos, tem efeito similar ao HGT. Por sua vez, duplicações gênicas e ILS não geram reticulações na árvore de espécies, mas podem fazer com que as árvores de genes apresentem comprimentos de ramos e até mesmo topologia diferentes. A partir do momento em que um gene duplica, suas diferentes cópias podem seguir caminhos evolutivos distintos. A possibilidade de duplicação seguida de extinção (DL – *gene duplication and loss*) de parte das linhagens gênicas é ainda mais grave, pois aumenta o risco de amostragem e utilização de cópias parálogas.

Diferente dos demais fenômenos, o ILS não envolve eventos de introdução de material genético no genoma, mas transcorre do processo natural de deriva das frequências gênicas nas populações ancestrais (Degnan & Rosenberg 2009). Se realizada, do presente em sentido ao passado, uma recapitulação do trajeto genealógico de qualquer par de cópias de um gene numa população, esse par necessariamente coalescerá (se encontrará) em seu ancestral mais recente (MRCA). O mesmo pode ser verificado para cópias amostradas de espécies distintas: descartada a possibilidade de introgressão, coalescerão próximo à especiação das linhagens que os contêm - em seu MRCA na população ancestral (**Figura 11a**). Em alguns casos, entretanto, quando o tamanho efetivo da população ancestral é grande o bastante para que cópias distintas de um gene coexistam nela por um longo período de tempo, ocorre a chamada coalescência profunda (**Figura 11b**), que, por si só, causa distorções entre os comprimentos de ramos (e tempos de divergência) da árvore do gene e os da árvore de espécies (Oliver, 2013). Dado que esse polimorfismo permaneça por gerações o bastante ao longo da história, a ponto de atravessar dois ou mais eventos de especiação, pode ocorrer, então, amostragem incompleta de linhagens, levando a discrepâncias entre suas topologias (**Figura 11c**).

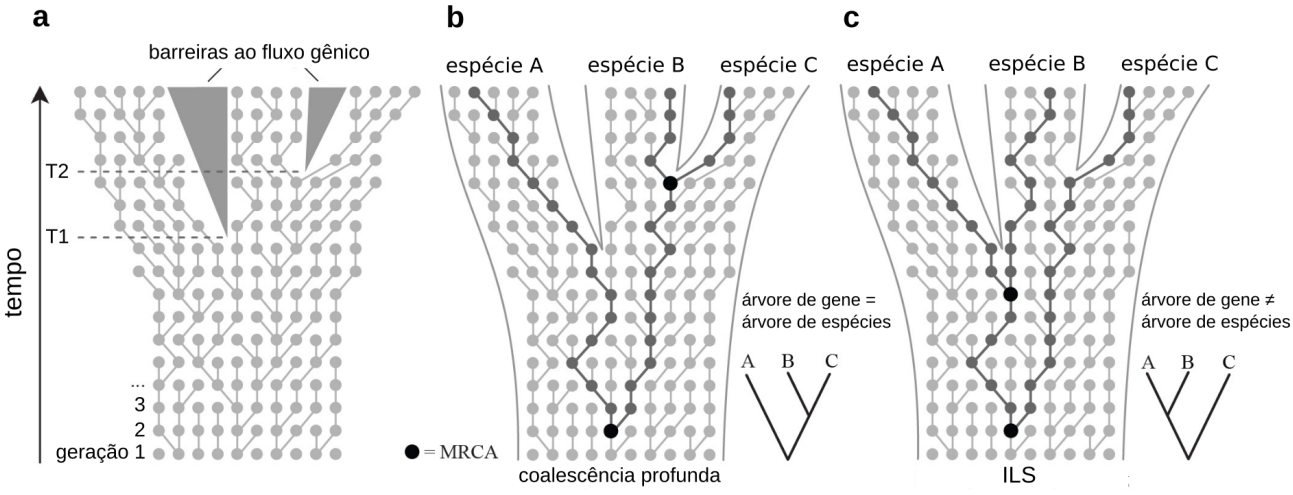


Figura 11: Coalescência de linhagens gênicas ao longo da árvore de espécies e as árvores de gene resultantes. (a) Na ausência de introgressão, cópias gênicas amostradas de diferentes espécies coalescem em gerações necessariamente anteriores ao surgimento de barreiras ao fluxo gênico, como T1 e T2; (b) em alguns casos, antecedendo o evento de especiação em um grande número de gerações (coalescência profunda). Quanto maior o tamanho efetivo (Ne) da população, por mais tempo as cópias tendem a coexistir nela. (b) Dado que esse tempo seja longo o bastante, a coexistência perdura através de um ou mais eventos de especiação, podendo causar ILS. Figura adaptada de Leliaert e colaboradores (2014).

Quando é de interesse do pesquisador inferir quais processos evolutivos ocorreram na história genômica do grupo estudado (e.g. duplicações e extinções de linhagens gênicas), a incongruência entre árvores de genes pode ser aproveitada como sinal desses processos

(Nakhleh 2013), porém, para o propósito de inferir uma árvore filogenética a partir de um alinhamento *multi-loci*, é de frequente interesse minimizar a heterogeneidade entre marcadores. Nesse sentido, é possível mitigar o impacto do ILS sobre as árvores de gene ao reduzir a amostragem taxonômica (Rosenberg, 2002). Como visto na seção anterior, entretanto, essa estratégia pode exacerbar o efeito da saturação de sequências (ao gerar ramos mais longos), em contrapartida.

Alternativamente, há métodos que buscam reconstruir a árvore de espécies por critérios de adequação às diferentes árvores de gene, se contrapondo aos que estimam uma única árvore diretamente dos genes concatenados em um bloco único. Apesar das primeiras abordagens do tipo terem buscado contemplar os efeitos de DL e HGT via parcimônia (Goodman et al. 1979; Maddison 1997), Maddison (1997) foi o primeiro a formalizar a aplicação de um modelo probabilístico (o coalescente, de Kingman 1982) para estimar árvores de espécies, levando em conta possíveis efeitos de ILS nas árvores de genes. Desde então, esse tipo de método, conhecido como coalescência multiespécies vem sendo cada vez mais aplicado.

A função proposta por Maddison (1997) para computar a verossimilhança de uma árvore de espécies incorpora tanto a probabilidade de cada árvore de gene possível, dada a árvore de espécies (pelo modelo coalescente), quanto as funções de verossimilhança de Felsenstein (1981) para cada árvore de gene (sob o modelo de substituição assumido):

$$L(ST) = \prod_{loci} \sum_{\substack{todas\ as\ GT \\ possíveis}} [p(locus|GT) \times p(GT|ST)]$$

Onde GT são as árvores de gene e ST, a árvore de espécies. O somatório das probabilidades de todas as árvores de gene possíveis para cada *locus* contemplaria toda a incerteza associada à estimação das árvores de genes, mas por ser virtualmente impeditivo computar a verossimilhança de cada árvore possível para um número moderado de táxons e múltiplos *loci*, a maioria dos métodos de coalescência dispensa esse procedimento e assume a árvore ML para cada *locus* como uma observação fixa, desconsiderando eventuais erros em sua estimação (Rannala & Yang 2017).

1.7 – O legado dos testes topológicos

Nesse cenário, em que não é possível acessar diretamente quanto erro estatístico é gerado ao assumir um determinado modelo evolutivo para um gene específico, testes topológicos ainda guardam grande potencial na era da Filogenômica, dado seu poder de

determinar quão significativas, na condição de estimativas, são as árvores de gene inferidas frente a outras topologias possíveis. Contudo, uma vez que a maioria dos testes de topologia exige razoável esforço computacional para conjuntos de dados com milhares de genes, esse potencial vem sendo cada vez menos aproveitado. Mesmo aproximações como o RELL podem não bastar para que esses testes sejam suficientemente rápidos para os conjuntos de dados do futuro.

Como visto na seção **1.5**, o único teste topológico que dispensa qualquer procedimento de replicação do alinhamento, além do recálculo das verossimilhanças das árvores comparadas, é a versão do KH que pressupõe normalidade da distribuição de Δln*L*s. Apesar de isso torná-lo um procedimento extremamente rápido, nele assume-se, também, que a variância da distribuição pode ser obtida da variância dos valores de Δln*L* calculados para os sítios do alinhamento original. Não há caso prático em que esse pressuposto adicional tenha sido comprovado e, naqueles (de frequente interesse) em que uma das árvores comparadas é a de ML, é possível que tal aproximação da variância discorde profundamente da esperada sob a hipótese nula do teste (igual adequação das árvores comparadas), condenando a estratégia ao viés de seleção inerente ao KH.

Intrigado por esse problema e inspirado na abordagem de Vuong (1989), busquei medir o efeito de diferentes fatores sobre a distribuição de Δln*L* entre árvores equidistantes da árvore verdadeira - uma abordagem até agora não explorada em filogenética. Com os resultados obtidos, desenvolvi um modelo de regressão experimental para aproximar intervalos de confiança para Δln*L*s, tendo como hipótese nula essa condição de equidistância. Ao lançar mão do mesmo pressuposto de normalidade do KH, mas utilizando um modelo de regressão para aproximar a variância de Δln*L*, esse novo teste via rapidez, ao dispensar procedimentos de replicação, sem a penalização de desempenho, pelo viés de seleção, que ocorre no KH.

Dado que os fatores analisados, por si só, expliquem suficientemente a variância de uma distribuição amostral de Δln*L*s, espera-se que o modelo de regressão apresente performance semelhante ao dos testes de hipótese tradicionais na determinação de um p-valor para o Δln*L* entre quaisquer duas árvores. Para avaliar essa performance, contudo, foi necessário aplicar o novo teste a dados empíricos. Com esse fim, pautei um dilema topológico em Mammalia, cuja dificuldade de resolução resiste aos anos apesar de ocorrer em um dos grupos mais bem representados em diversidade de genomas sequenciados.

1.8 – A raiz dos placentários

Apesar das relações filogenéticas entre as infraclasses de Mammalia (Monotremata, Marsupialia e Placentalia) mostrarem-se bem estabelecidas por múltiplas evidências (Luo et al. 2001; Phillips & Penny 2003; Meredith et al. 2011), as resoluções filogenéticas para relações entre algumas superordens e ordens dentro de Marsupialia e Placentalia permanecem controversas. A difícil resolução dos parentescos mais antigos nestes grupos tem afetado mais notadamente a raiz (primeiro evento de especiação) da árvore dos placentários. Diferentes conjuntos de dados e métodos de inferência tem recuperado relações discrepantes entre seus grandes grupos.

Os placentários podem ser divididos em quatro superordens, das quais Xenarthra agrupa tatus, tamanduás e preguiças; Afrotheria, os peixes-boi, toupeiras-douradas e elefantes; Laurasiatheria conta com canídeos, equinos, cetáceos e morcegos, entre outros e, por fim, Euarchontoglires inclui todos os primatas e roedores. Até hoje, três hipóteses ainda são consideradas fortes candidatas a topologia verdadeira em relação à raiz de Placentalia: (1) Xenarthra como grupo irmão do restante da diversidade placentária (O'Leary et al. 2013), (2) Afrotheria ocupando esta posição (Murphy et al. 2001; McCormack et al. 2012; Romiguier et al. 2013) ou (3) um clado contendo Afrotheria e Xenarthra (chamado, nesses casos, de Atlantogenata) como grupo irmão de Boreoeutheria (que agrupa Euarchontoglires e Laurasiatheria e é recuperado em todos os três cenários) (Meredith et al. 2011; Morgan et al. 2013) (**Figura 12**).

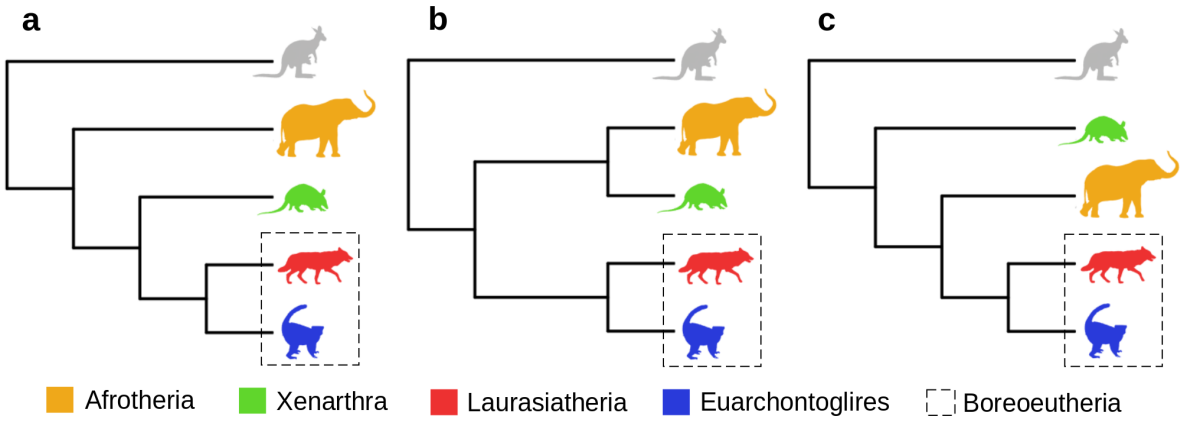


Figura 12: **Três resoluções mais aceitas para o posicionamento da raiz de Placentalia**: (a) Entre Afrotheria e (Xenarthra + Boreoeutheria); (b) entre Atlantogenata e Boreoeutheria e (c) entre Xenarthra e (Afrotheria + Boreoeutheria).

Essa discrepância de resoluções já foi associada a uma série de fatores, dos quais alguns já apresentados nessa introdução; como saturação de sítios (Phillips et al. 2006;

Nilsson et al. 2010), vieses de composição - quando os eventos de substituição entre certas bases acumulam e saturam mais rapidamente que entre outras - (Foster & Hickey 1999; Phillips et al. 2006), variação das taxas de substituição entre diferentes linhagens (Li et al. 1987) e, nesse contexto, especificação inadequada de modelos evolutivos (Philippe et al. 2011). Buscando reduzir o impacto destes fatores na recuperação da filogenia de Mammalia, alguns autores sofisticaram suas análises, seja assumindo modelos de substituição separados para cada gene (Nishihara et al. 2007), aplicando métodos de coalescência multiespécies (Song et al. 2012), aumentando o número de parâmetros no modelo de substituição utilizado (Morgan et al. 2013) ou priorizando marcadores ricos em AT para contornar um viés composicional (Romiguier et al. 2015). Mesmo assim, essas aproximações recuperam árvores discrepantes entre si, em alguns casos discordando até mesmo na posição da raiz.

Visto que há milhares de genes ortólogos sequenciados e publicamente disponibilizados para centenas de espécies de placentários e que suas antigas relações permanecem desafiadoras apesar dos avanços analíticos, a filogenia do grupo foi a escolha natural como estudo de caso deste trabalho. Voltando a esse problema, pude sondar não somente a utilidade de testes topológicos na era da Filogenômica, como a significância do sinal filogenético em múltiplos genes para as diferentes hipóteses da raiz de Placentalia.

2 – Objetivos

2.1 – Geral

Com o desenvolvimento deste trabalho, busco investigar o comportamento da estatística $\Delta\ln L$ entre pares de árvores que sejam explicações igualmente razoáveis do processo evolutivo de um gene qualquer, assim como a possibilidade de prever sua dispersão em diferentes cenários. Com os resultados obtidos, desejo aprimorar a relação custo computacional vs. poder de testes topológicos que implementem essa estatística e, assim, sua conveniência para a resolução de grandes dilemas topológicos no contexto da Filogenômica.

2.2 – Específicos

(1) Avaliar a resposta da distribuição de $\Delta\ln L$ à variação de diferentes componentes no alinhamento e nas topologias comparadas, enquanto equidistantes da árvore verdadeira, no espaço topológico.

(2) Criar um modelo de regressão que seja capaz de prever, com mínimo custo computacional, o intervalo de confiança da distribuição de $\Delta\ln L$ entre quaisquer duas árvores na condição de equidistância.

(3) Utilizar o modelo de regressão e testes de hipótese tradicionais para verificar o viés de múltiplos marcadores ortólogos para Mammalia, entre as estimativas pontuais de ML para suas árvores de genes e as hipóteses alternativas de posicionamento da raiz de Placentalia.

(4) Comparar o performance do novo teste aos testes topológicos tradicionais.

(5) Verificar, à luz dos testes topológicos e em comparação a estimativas pontuais por diferentes abordagens, a contribuição de cada procedimento para a resolução da raiz de Placentalia.

3 – Metodologia

3.1 – Análises da distribuição de Δln*L*

Para analisar as propriedades da estatística Δln*L*, reproduzi sua distribuição nos casos em que as árvores comparadas estão equidistantes da árvore verdadeira, ou seja, da árvore que representa com exatidão o processo que gerou os dados. Como é impossível determinar qual é a exata árvore que gerou um conjunto de dados empíricos, foi necessário obter valores de Δln*L* a partir de alinhamentos gerados por simulação de sequências. As topologias entre as quais foram calculados os Δln*L*, por sua vez, foram obtidas através de um procedimento de rearranjo de cada árvore verdadeira. Ao longo de todo esse processo, quatro fatores específicos foram controladamente variados para verificação de sua influência sobre a distribuição de Δln*L*: o número de ramos nas árvores, a distância entre as árvores no espaço topológico, a extensão e a proporção de *gaps* dos alinhamentos simulados.

3.1.1 – Parametrização das simulações e topologias verdadeiras

Em geral, simulações de sequências de DNA se dão pelos mesmos modelos de Markov aplicados na estimação de árvores filogenéticas. Nesse caso, sua função é revertida para determinar, ao invés de inferir, os processos de substituição que ocorreriam ao longo de uma árvore fixada *a priori,* gerando um conjunto de sequências diferentes, mas aparentadas, ao seu fim. Para essas sequências, as árvores fixadas são as árvores verdadeiras com as quais se deseja trabalhar aqui; portanto, primeiro foi necessário delimitar quais seriam as árvores, o modelo de substituição e respectivos parâmetros a guiar as simulações.

Os parâmetros de substituição e a topologia da árvore verdadeira podem ser definidos arbitraria ou "empiricamente". No primeiro caso, os parâmetros são fixos de acordo com interesses específicos do pesquisador; por exemplo, quando se investiga a influência do valor real dos parâmetros na acurácia de métodos de inferência (ver Schwartz & Muller 2010). No segundo, são estimados de dados empíricos, independentes das análises a serem realizadas posteriormente. Entendendo que essa última estratégia é mais realista biologicamente, por se basear na estimativa de um processo evolutivo que realmente ocorreu, optei por empregá-la. Todos os parâmetros do modelo de substituição e as árvores utilizadas em minhas simulações foram estimados por máxima verossimilhança de um concatenado de 26 marcadores nucleares (36.548 *bp*) para 290 espécies de mamíferos. Esse alinhamento foi antes utilizado no trabalho

de Foley, Springer e Teeling (2016) - que debate diferentes conflitos topológicos na árvore dos placentários - e recuperado para minha utilização através da plataforma Dryad.

O modelo de Markov escolhido para inferência dos parâmetros evolutivos do concatenado foi o GTR (*general time reversible* – Tavaré 1986), por permitir a livre variação de todas as taxas de substituição e frequências de equilíbrio. Sendo o caso geral de todos os demais modelos reversíveis de substituição de DNA, o GTR frequentemente explica o processo de substituição em dados reais mais adequadamente que modelos menos complexos (Arenas 2015). Dada a extensão do concatenado, assumi, também, taxas de substituição heterogêneas ao longo dos sítios. Para determinar a probabilidade de diferentes taxas, empreguei uma distribuição gama com oito categorias. Apesar de quatro categorias gama mais uma de sítios invariáveis (Waddel & Steel 1997) serem uma aplicação mais comum, utilizando oito categorias espero contemplar tanto sítios de variação muito lenta quanto mais acelerada. Com isso, é possível dispensar uma - possivelmente pouco realista - categoria de sítios que não mudam em escala macro-evolutiva, além de evitar lentidão com os problemas de otimização consequentes da estimação do parâmetro alfa (que determina a forma da distribuição gama) e da proporção de invariáveis ao mesmo tempo.

Nesses termos, comecei por inferir uma árvore incluindo todos os 290 táxons do concatenado, utilizando o programa IQ-tree (Nguyen et al. 2014). Desta árvore, foram excluídos os táxons que apresentaram distância filogenética em relação a seu táxon irmão (soma dos comprimentos de seus ramos terminais) inferior a 0.03 (substituições por sítio – s/s). Esse procedimento se deu para evitar a simulação de alinhamentos com sequências idênticas mais tarde. Por meio de um *script* em Perl, a cada dupla de táxons irmãos que se enquadrasse no critério, aquele com menor comprimento de ramo foi excluído da árvore inferida e, assim progressivamente, até que todos os ramos terminais e irmãos somassem mais de 0,03 s/s. Em seguida, a sequências dos 180 táxons remanescentes foram selecionadas para obter uma versão reduzida do alinhamento original. Adicionalmente, foram feitas sub-amostragens aleatórias desses 180 táxons para obter outros três alinhamentos com 130, 80 e 30 táxons cada.

Cada um dos quatro novos alinhamentos, por sua vez, foi utilizado para estimar, também sob o modelo GTR e 8 categorias gama, as árvores e parâmetros de substituição de ML que governaram as simulações dos alinhamentos analisados posteriormente. Entre os parâmetros estimados incluem-se taxas de substituição; frequências das bases e o parâmetro alfa. A partir daqui, essas quatro árvores de diferentes tamanhos são referidas ‘árvores verdadeiras’, bem como suas topologias totalmente resolvidas. Seus diferentes números de

táxons permitiram, depois, avaliar o efeito do número de ramos na topologia sobre a distribuição de $\Delta \ln L$.

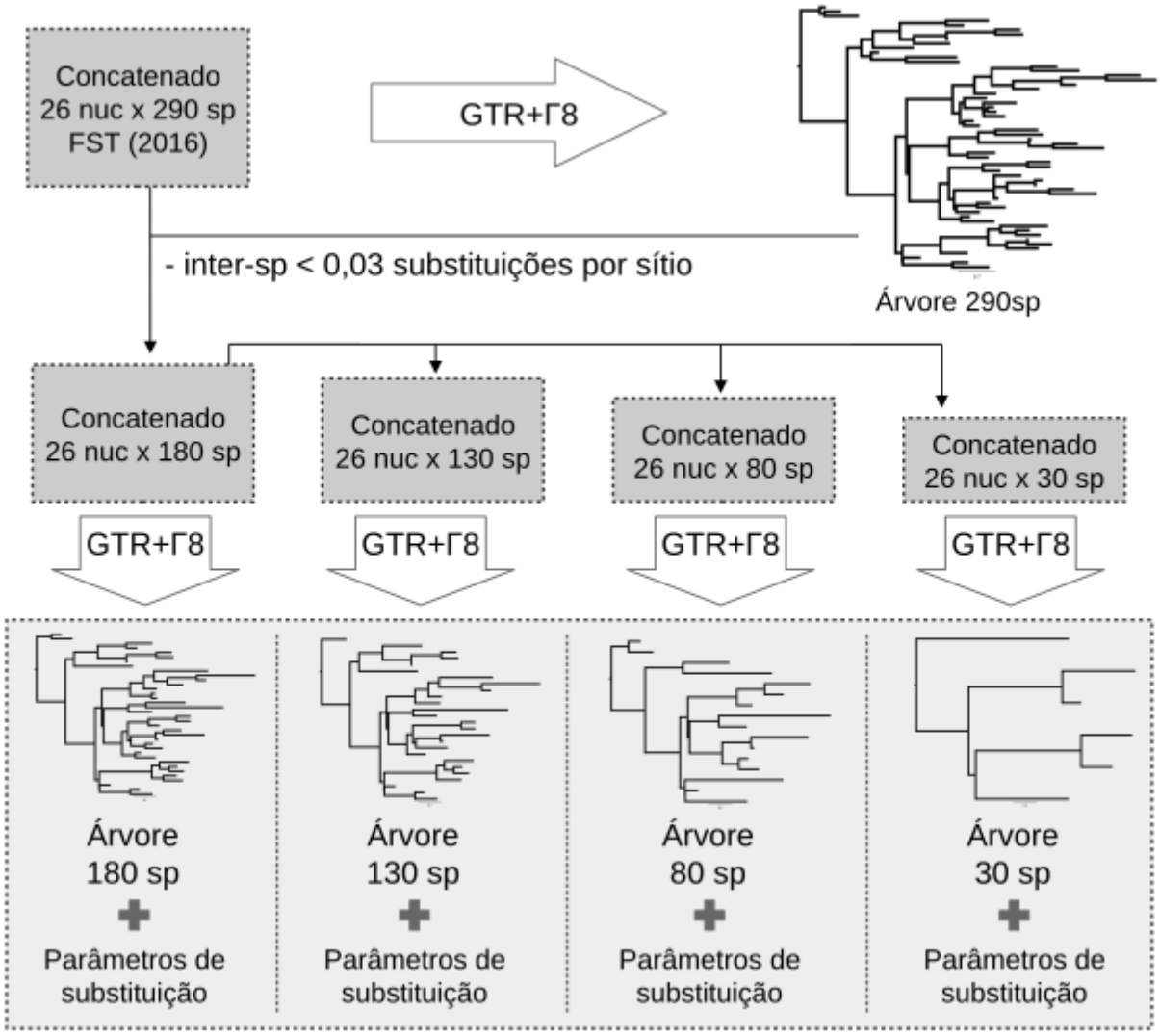


Figura 13: Geração das árvores e parâmetros de substituição que guiaram as simulações dos dados. Cada seta branca corresponde a uma inferência filogenética no programa IQ-tree com a matriz de substituição e número de categorias gama (Γ) indicados. As setas delgadas indicam a amostragem de sequências do alinhamento de Foley, Springer e Teeling (2016) e da versão reduzida com 180 espécies, respectivamente. No primeiro caso, a amostragem se deu pelo critério de distância filogenética entre os taxons irmãos (> 0,03 s/s) na árvore de 290 espécies; no segundo, se deu aleatoriamente para a obtenção de alinhamentos (e árvores) com diferentes números de taxons. O quadro em cinza claro destaca os sets de árvores e parâmetros de substituição verdadeiros, que foram, depois, utilizados na simulação de múltiplos alinhamentos. Todas as árvores representadas aqui são ilustrativas, não correspondendo às árvores estimadas.

3.1.2 – Obtenção das topologias alternativas

De forma a obter as topologias alternativas, das quais foram posteriormente computados os valores de verossimilhança (e a diferença entre eles - $\Delta \ln L$) utilizando os

alinhamentos simulados, cada uma das quatro topologias verdadeiras foi rearranjada independente e aleatoriamente pelo método de *nearest neighbor interchange* (NNI – Robinson 1971) (**Figura 14**). Nesse procedimento, um dos ramos internos da árvore é substituído por uma resolução outra, de forma a obter uma topologia diferente. Mais passos de NNI geram árvores progressivamente mais diferentes da original. Entre outros métodos de rearranjo, como o SPR (*subtree prunning and regrafting*) e o TBR (*tree bisection and reconnection*), o NNI é o que faz alterações menos radicais na árvore, sendo necessário um número maior de passos para alcançar topologias mais distantes, mas o que permite maior controle da distância percorrida no espaço topológico (ver St. John 2016).

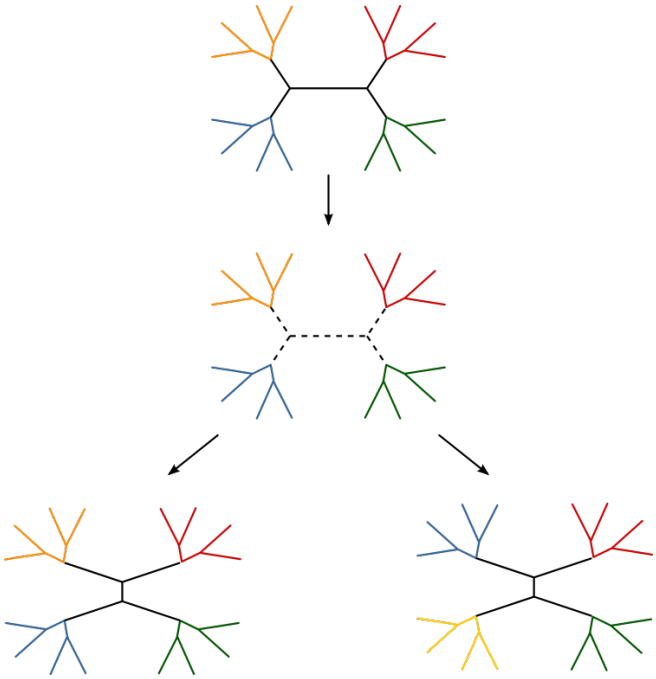


Figura 14: Rearranjo por NNI de uma árvore com 16 táxons. Um de seus ramos internos é selecionado aleatoriamente e substituído, junto a seus vizinhos mais próximos, por uma das duas outras resoluções possíveis para as relações entre as sub-árvores que separam. Figura retirada de commons.wikimedia.org/.

O rearranjo por NNI de cada uma das topologias foi realizado múltiplas vezes e com quantidades diversas de passos, de forma a obter pares de árvores que tivessem (1) diferentes distâncias topológicas entre si e que, ao mesmo tempo, as árvores em cada par apresentassem (2) a mesma distância topológica em relação à árvore verdadeira da qual foram rearranjadas. A primeira conjuntura me permitiu investigar a influência da distância entre as árvores comparadas sobre a distribuição de $\Delta \ln L$. Já a segunda, garantiu a mesma condição ensejada

na abordagem de Vuong (1989): a equidistância entre os modelos probabilísticos comparados e o modelo verdadeiro.

No entanto, ao passo que Vuong utilizou a divergência de Kullback-Leibler (Kullback & Leibler 1951) para abordar a distância entre modelos, aproveitamos uma das métricas desenvolvidas nas últimas décadas para distâncias entre árvores filogenéticas: a *branch score distance* (BSD), de Kuhner e Felsenstein (1994). Entre duas árvores com o mesmo conjunto de táxons, a BSD é o somatório das diferenças entre os comprimentos de seus ramos que determinam cada partição possível desse conjunto de táxons. Deste modo, ela é ponderada não somente pelas diferenças entre as topologias, mas também pelas diferenças entre os comprimentos dos ramos que estão presentes em ambas as árvores.

Todo o procedimento de rearranjo das quatro topologias verdadeiras e seleção das árvores alternativas foi realizado através de um *script* que produzi em *R* (R Core Team 2016), envolvendo funções dos pacotes *ape* e *phangorn* (**Suplementar 1**). Com o *script*, selecionei trios de árvores alternativas que fossem equidistantes da árvore verdadeira da qual foram rearranjadas. Dado que com duas árvores alternativas tem-se apenas um par, mas, entre três, há três pares possíveis, a seleção em trios me permitiu aumentar a razão entre a quantidade de $\Delta\ln L$s calculados (um para cada par) e o tempo necessário para calcular as verossimilhanças de todas as árvores alternativas. Cada trio de árvores foi selecionado pelas diferentes distâncias entre elas, que variaram em torno de 0.05 até 0.10, 0.15, 0.20, 0.25 ou 0.30 BSD. Assim, obtive 3 trios para cada distância e de cada árvore verdadeira (original), somando um total de 216 árvores alternativas. Nessa etapa, as distâncias entre cada trio e sua árvore verdadeira só foi levada em conta para garantir a condição de equidistância (**Figura 15**).

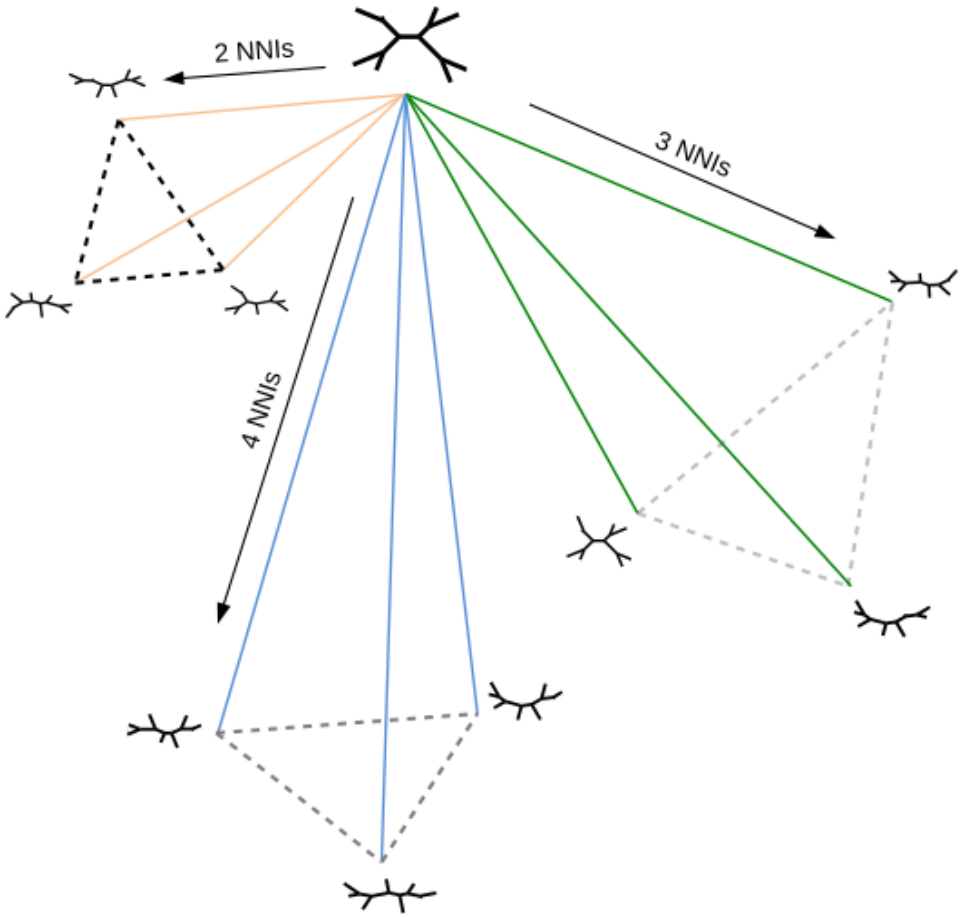


Figura 15: Representação com árvores não enraizadas, de 8 táxons, do esquema de rearranjos realizado com cada árvore verdadeira, em uma abstração tridimensional do espaço topológico. Cada seta apresenta o número de passos de NNI realizados para obter um novo trio de árvores. Das linhas que se interceptam, as de mesma cor e padrão representam distâncias iguais. As sólidas representam as distâncias entre cada árvore alternativa e a verdadeira; cada cor remetendo a um *set* de rearranjos independente. As linhas tracejadas, por sua vez, representam as distâncias entre as árvores alternativas de cada trio - quanto mais claras, maiores. Maior distância da árvore verdadeira não necessariamente implica maior distância entre as árvores alternativas.

3.1.3 – Simulação dos alinhamentos

Uma vez que viso explorar, também, o efeito da proporção de *gaps* sobre a distribuição de $\Delta \ln L$, escolhi o programa INDELible (Fletcher & Yang 2009) para simular os alinhamentos, porque implementa um modelo de inserção e um de deleção - paralelamente ao modelo de substituição - ao longo da geração das sequências. Esses modelos amostram, de uma determinada distribuição de probabilidades, o tamanho do vetor de nucleotídeos a ser inserido ou deletado das sequências (*indel*), com frequência dada por uma taxa de inserção (ou deleção) especificada. Como nenhum programa de inferência filogenética conhecido por

mim implementa esse tipo de modelo, foi necessário definir seus parâmetros de forma parcialmente arbitrária.

Fixei o mesmo modelo para os eventos de inserção e deleção, amostrando de uma distribuição de Pascal (binomial negativa) com tamanho 1 e probabilidade 0,5 (parâmetros 'r' e 'q' no INDELible). Considerei razoável utilizar esses parâmetros porque lido com os sítios dos alinhamentos como observações independentes, não importando, portanto, a extensão de cada *indel* gerado. Contudo, como era de interesse produzir alinhamentos com diferentes proporções de *gap*, foi necessário realizar simulações preliminares com as quatro árvores verdadeiras, no intuito de determinar a taxa de *indel* mais apropriada para cada combinação de tamanho da árvore geradora e proporção de *gaps* desejada. Assim, foi regredido o modelo

$$G \sim 0{,}01669 \times L + 9{,}69185 \times I - 0{,}11545$$

, onde I é a taxa de *indel*, L é o comprimento da árvore (soma dos comprimentos de todos os ramos) e G é a proporção de *gaps* resultante.

Utilizando as taxas de *indel* previstas por esse modelo, as árvores verdadeiras e os parâmetros de substituição estimados com elas, simulei e triei múltiplos alinhamentos, que variaram em proporção de *gaps* de 0%, a aproximadamente 10, 20, 30 ou 40%; e em extensões de 100, 200, 300, 400, 500, 750, 1000, 1500 ou 2000 bp. A variação na extensão dos alinhamentos, mais uma vez, teve intuito de contemplar o efeito deste outro fator na distribuição de $\Delta \ln L$. Como os eventos de inserção e deleção ao longo das simulações aumentam a extensão dos alinhamentos para além do tamanho requisitado, utilizei, em conjunto com o INDELible, um *script* em *Perl* para sortear um *set* de sítios de tamanho exato ao desejado, de cada alinhamento gerado.

3.1.4 – Otimização de parâmetros e cálculo dos valores de $\Delta \ln L$

Em sequência, para pontuar as topologias obtidas de acordo com os dados gerados, cada réplica de alinhamento simulada foi utilizada para otimizar os parâmetros tanto das topologias alternativas quanto das verdadeiras e obter seus valores de verossimilhança. Esse procedimento se assemelha à busca completa pela árvore de ML, mas, nele, encontra-se a combinação de comprimentos de ramos e de parâmetros no modelo de substituição que maximizam a verossimilhança da árvore inicial (dada pelo usuário) sem alterar sua topologia; ou seja, sem utilizar nenhuma heurística para transpassar o espaço topológico (como NNI,

SPR, etc.). Para otimização de todas as árvores foi utilizado o mesmo modelo de substituição que gerou os alinhamentos (GTR com 8 categorias gama), no mesmo programa IQ-tree.

Sabendo, entretanto, que a métrica BSD é sensível aos comprimentos de ramo das árvores comparadas, foi necessário, após a otimização de parâmetros, triar os pares de árvores que se mantiveram virtualmente equidistantes (com diferença inferior a 0,01 no BSD) em relação à árvore verdadeira da qual foram rearranjados. Dos pares que continuaram satisfazendo ao critério, um total de 2.989.720 valores de diferença entre suas verossimilhanças ($\Delta \ln L$s) foi obtido. Como foram utilizados alinhamentos de diferentes extensões para obter esses valores, foi necessário, também, padronizá-los para torna-los comparáveis. Assim, qualquer menção a '$\Delta \ln L$', a partir daqui, trata da diferença média entre as log-verossimilhanças por sítio de duas árvores; ou seja, da diferença entre as log-verossimilhanças totais de cada uma dividida pelo número de sítios no alinhamento pela qual foram computadas ($[\ln L1 - \ln L2] / N$).

Da etapa de simulação dos alinhamentos até a seleção dos pares equidistantes, todos os procedimentos foram automatizados através de um *pipeline* em Shell (**Suplementar 2**). Implementando linhas de comando do INDELible e do IQ-tree, além de scripts em perl e em R, esse *pipeline* simulou múltiplos alinhamentos para cada combinação de árvore verdadeira, número de sítios e proporção de *gaps*. Dado cada alinhamento simulado, otimizou comprimentos de ramo e calculou verossimilhanças fixando ora a topologia de cada árvore verdadeira, ora de suas árvores alternativas com diferentes distâncias entre si (**Figura 16**). Ao mesmo tempo, compilou uma longa tabela de 1,5GB que inclui identificação de cada árvore otimizada, seu número de táxons, distância topológica em relação a seu par e a sua árvore verdadeira, extensão e proporção de *gaps* do alinhamento correspondente, estimativa do alfa da distribuição gama, entre outros dados; facilitando, assim, a manipulação dos resultados.

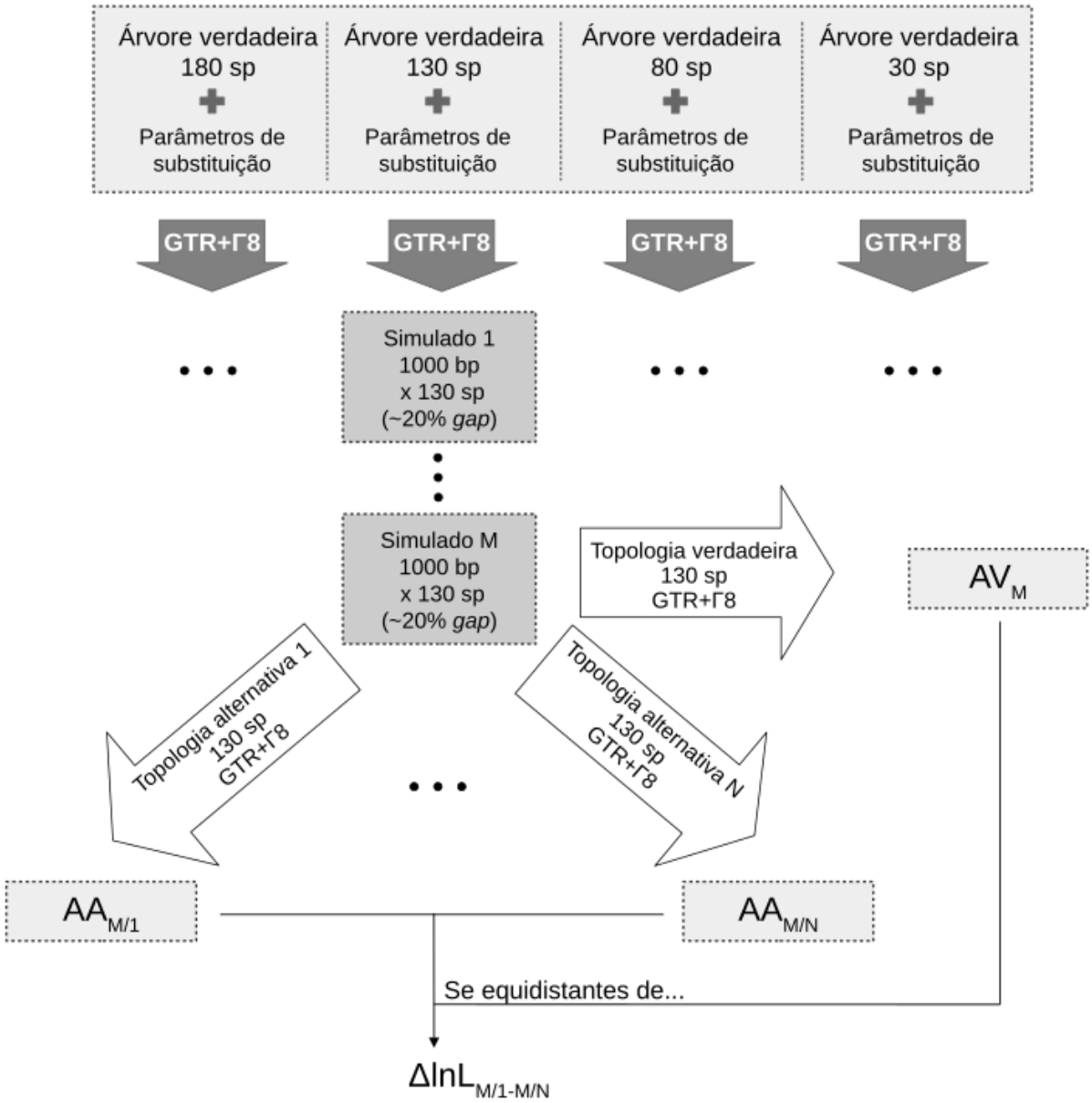


Figura 16: Esquema de simulações dos alinhamentos e otimizações das árvores verdadeiras e alternativas. Após parametrização das simulações com as árvores, frequências de bases e taxas de substituição estimados das sequências de Foley, Springer e Teeling (2016), cada alinhamento simulado serviu para estimar os comprimentos de ramos na topologia da árvore que o gerou e nas topologias de cada uma das árvores alternativas de mesmo número de táxons (e com diferentes distâncias entre si). Cada seta em cinza escuro indica a simulação de múltiplos alinhamentos sob a mesma árvore verdadeira, enquanto as brancas exemplificam estimações de comprimentos de ramos em topologias fixas. AV_i e $AA_{i/j}$ são, respectivamente, a árvore verdadeira e a alternativa j, otimizadas com o alinhamento i. Apesar do exemplo mostrar apenas otimizações para o último de um set com M alinhamentos, o procedimento se repete para todos os M e, também, para os de outros sets de alinhamentos não representados na figura, com diferentes dimensões e proporções de gaps.

As análises dos dados resultantes foram todas realizadas em R. Primeiramente, foi feita uma avaliação gráfica da mudança na dispersão dos valores de Δln*L* mediante variação

de cada fator que passou por algum controle (parcial para as BSDs entre árvores alternativas, alteradas após otimização dos ramos), além de outras variáveis não controladas. Essa etapa foi necessária para determinar se a condição de equidistância garante equidade de adequação entre as árvores comparadas, o que se refletiria numa dispersão simétrica dos valores de $\Delta\ln L$ em torno de 0, independente da variação em demais fatores. Além disso, foi possível determinar, também, se e como cada variável afeta a difusão dos valores de $\Delta\ln L$.

Por fim, para cada combinação possível de diferentes valores das variáveis que demonstraram alguma influência sobre a difusão de $\Delta\ln L$, foram amostrados aleatoriamente 200 $\Delta\ln L$s. Para isso, as observações com diferentes distâncias BSD - única variável de distribuição contínua entre as selecionadas – tiveram de ser separadas em seis categorias descontínuas de distâncias, indo de aproximadamente 0,02 até 0,3 BSD e com largura de 0,03 cada. Assim, as 1.080 amostras de $\Delta\ln L$ obtidas formaram distribuições distintas, das quais obtive os desvios padrões (SDs – *standard deviations*) para avaliar sua correlação com cada uma das variáveis selecionadas.

3.1.5 – Regressão múltipla dos valores de Δlni

Buscando determinar se as variáveis apresentavam relação com a difusão das distribuições de $\Delta\ln L$, como também definir quão predominante seria a influência de cada um sobre esta difusão, realizei análises de regressão linear entre elas e os SDs das distribuições. Enquanto a regressão linear simples permitiu estimar um coeficiente de regressão (i.e. o efeito) de cada variável independente sobre a dependente (o SD de $\Delta\ln L$) separadamente, a regressão linear múltipla (MLR – m*ultiple linear regression*) foi ainda mais útil, por possibilitar a estimação dos múltiplos efeitos das diferentes variáveis independentes sobre a dependente, ou seja, a contribuição relativa de cada uma para a variação do SD.

Os coeficientes de regressão são estimados sobre um modelo proposto *a priori*, que, na MLR, segue o formato '$Y = ß_0 + ß_1X_1 + ß_2X_2 + ß_3X_3 + E$' para três variáveis independentes (X_i) e uma dependente (Y), por exemplo. Neste caso, $ß_{1\text{-}3}$ são os coeficientes de regressão de cada X_i sobre Y, a serem estimados; $ß_0$ é o intercepto do modelo (valor que Y assume quando todas as X_i são nulas), também a ser estimado e E é o total dos erros associados à estimação de cada parâmetro ß. Por sua vez, E é quantificado através dos resíduos (diferença entre os valores de Y observados e os preditos pelo modelo regredido para Y – ou Y^).

No entanto, a validade de qualquer modelo estimado via MLR depende, antes de tudo, da satisfação de quatro pressupostos: (1) ausência de multicolinearidade (i.e. independência

mútua entre as variáveis independentes), (2) linearidade da relação entre as variáveis independentes e a dependente, (3) homocedasticidade (dispersão homogênea) dos resíduos e (4) normalidade multivariada das variáveis independentes (Hair et al. 2009). O fato de que cada distribuição de $\Delta\ln L$ amostrada referiu-se a uma combinação específica de valores (ou categorias de valores) das variáveis selecionadas garantiu a independência entre elas em meu experimento e, portanto, a satisfação do primeiro pressuposto. Entretanto, outros procedimentos *a posteriori*, focados principalmente em análises gráficas, foram necessários para garantir a coincidência de todas as condições.

Para atestar a linearidade entre variáveis independentes e a dependente realizei dois tipos de abordagens gráficas: análise dos *boxplots* dos SDs das distribuições de $\Delta\ln L$ (variável dependente) que se encaixavam em cada categoria de cada variável independente e análise dos *plots* dos resíduos ao longo dos valores previstos pelo modelo de MLR para a variável independente (**Figura 17**). Nos casos em que a primeira abordagem indicou relação monotônica (consistentemente positiva ou negativa), porém não-linear com o SD, o fator e/ou o SD foi transformado (e.g. logaritmizado, invertido ou elevado ao quadrado) até atingir-se linearidade; optando-se pela exclusão da variável apenas na ocasião de nenhuma transformação simples aprimorar a linearidade.

Já a segunda abordagem não só permitiu confirmar a linearidade das relações entre variáveis independentes e dependente, como também verificar a terceira condição da MLR (homocedasticidade dos resíduos). Ao passo que a violação do primeiro leva a um padrão não-linear da dispersão dos resíduos ao longo dos valores previstos (**Figura 17a**), a violação da homocedasticidade (i.e. heterocedasticidade) ocorre quando há variação do seu desvio padrão (**Figura 17b**). Pode ocorrer, inclusive, coexistência dessas distorções (**Figura 17c**). Transformações adicionais das variáveis e do SD de $\Delta\ln L$, intercaladas com novas regressões do modelo, garantiram máxima aproximação do padrão linear e homocedástico (**Figura 17d**).

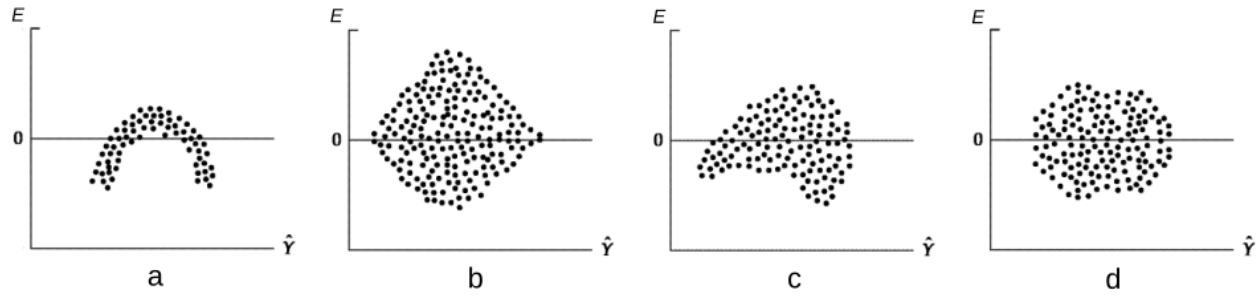


Figura 17: Exemplos de padrões em análises de resíduos. A violação de diferentes pressupostos do MLR pode levar a diversos tipos de ditorções no padrão de dispersão dos resíduos (E) ao longo dos valores previstos (Y^). Por exemplo, (a) relação não-linear entre variáveis depependentes e

independente, (b) heterocedasticidade dos resíduos ou (c) heterocedasticidade e não-linearidade concomitantemente; sendo desejável atingir-se (d) dispersão aproximadamente linear e homocedástica. Gráficos adaptados de Hair et al. 2009.

Por fim, a normalidade multivariada das variáveis independentes foi verificada avaliando se esses mesmos resíduos apresentavam distribuição normal. Na mesma abordagem que intercalou transformações com a regressão do modelo, garanti essa última condição através da avaliação do histograma dos resíduos e do *plot* dos quantis da distribuição de resíduos contra os quantis da função normal padrão (também conhecido como *plot* quantil-quantil) (**Figura 18**). Essas análises são mais apropriadas que testes de normalidade na presença de um grande número de observações, porque, nesses casos, os testes tendem a rejeitar a hipótese de normalidade sob qualquer pequeno desvio da condição, geralmente inflando sua taxa de erro do tipo I (Ghasemi & Zahediasl 2012).

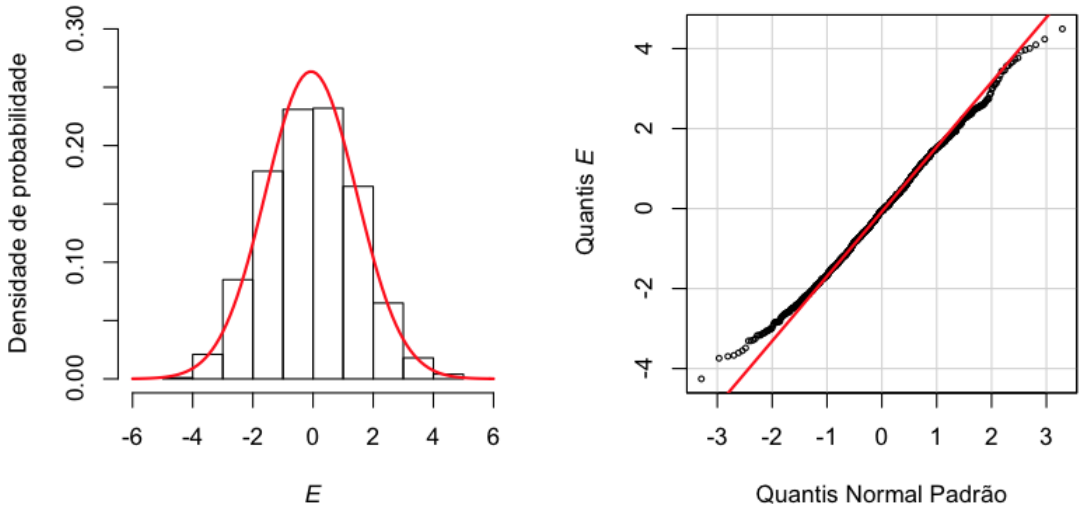


Figura 18: Verificação gráfica da normalidade de uma distribuição de resíduos. A normalidade de qualquer distribuição pode ser averiguada comparando seu histograma à curva de densidade de probabilidade normal para mesma média e desvio padrão (à esquerda) ou plotando seus quantis contra os quantis de uma normal padrão (á direita) e verificando se foge consideravelmente da correlação esperada com uma normal de mesma média e desvio padrão (linha diagonal). Os "resíduos" utilizados na composição desses exemplos advém de uma distribuição normal gerada aleatoriamente no R, com média 0 e desvio padrão 1,5.

Após adição sucessiva das variáveis que mostraram maior até menor relevância ao R^2 (coeficiente de determinação) do modelo, a combinação que permitiu melhor acomodação aos quatro pressupostos da MLR foi selecionada. O modelo resultante foi então aplicado para previsão do desvio padrão esperado para $\Delta \ln L$ na condição de equidistância entre as árvores

comparadas e a árvore que gerou o alinhamento. Esse propósito foi ensaiado pela implementação do modelo em um novo teste topológico, que nomeei *Equidistant Delta*.

3.2 – Equidistant Delta

Ao invés de testar a hipótese de que uma ou outra árvore seja a verdadeira, o *Equidistant Delta* (ED) testa a hipótese de que as duas árvores sendo comparadas estão equidistantes - no espaço topológico - da árvore que verdadeiramente representa o processo de evolução do alinhamento analisado. Baseando-se no teorema do limite central (TLC), esse teste aproxima um IC para os valores de Δln*L* possíveis entre as duas árvores, na ocasião da hipótese nula ser verdadeira.

O TLC postula que, obtidas múltiplas amostras de uma dada variável aleatória e independentemente distribuída, o conjunto das médias dessas amostras se aproximará de uma distribuição normal, mesmo que a distribuição subjacente (da variável em si) não seja normal (Feller, 1945). Da mesma forma, é esperado que, dados vários alinhamentos que evoluíram sob o mesmo processo, os valores de ΔlnL entre duas árvores que representem esse processo de maneira igualmente adequada (i.e. equidistantes da árvore verdadeira) apresentem, também, distribuição aproximadamente normal.

O teste de KH se baseia no mesmo princípio quando assume a normalidade da distribuição amostral de Δln*L*, mas, além disso, assume que a variância dessa distribuição equivale à da distribuição de Δln*L* por sítio do alinhamento original. Já no teste que proponho aqui, a variância esperada para Δln*L* quando a hipótese nula é verdadeira, é, por definição, o quadrado do SD previsto pelo modelo de regressão. Da mesma forma, os limites à esquerda e à direta de um intervalo de confiança (IC) para Δln*L* podem ser obtidos multiplicando o SD previsto pelos quantis 2,5% e 97,5% da normal padrão. Por sua vez, esse IC delimitaria os valores que o Δln*L* médio entre duas árvores poderia apresentar sem rejeitar a hipótese de que ambas sejam equidistantes da árvore verdadeira (**Figura 9**).

Desde que, com dados empíricos e nas condições da hipótese nula, as variáveis contempladas pelo modelo sejam suficientes para explicar a maior parte da variação no SD da distribuição de Δln*L* e que essa distribuição tenda de fato à normalidade e tenha média 0, é esperado que o ED controle razoavelmente bem sua taxa de erro do tipo I; mesmo que uma das árvores comparadas seja a de máxima verossimilhança. Além disso, a presunção de normalidade e o modelo preditor dispensam a necessidade de se replicar os valores de Δln*L* – via *bootstrap* ou RELL - no intento de obter um IC; procedimento que permanece indispensável nos demais testes topológicos, além do KH.

3.3 – Análises dos dados empíricos

Para avaliar o performance do teste proposto e, através dele, mapear o viés filogenético de dados empíricos em relação ao dilema da raiz de Placentalia, realizei análises de múltiplos marcadores ortólogos para o grupo. Essas análises envolveram a estimação de árvores de gene por máxima verossimilhança e execução de múltiplos testes de hipótese entre as árvores de ML para cada gene e cada alternativa para o posicionamento da raiz. A etapa de testes topológicos não se limitou à aplicação do modelo desenvolvido, incluindo outros testes tradicionalmente utilizados.

3.3.1 – Amostragem das sequências e taxonomia

Os dados empíricos analisados foram sequências de DNA codificante (*CDSs*) obtidas no banco de dados OrthoMaM v.9 (Orthologous Mammalian Markers - Douzery et al. 2014). Alinhado com a base de dados Ensembl (Aken et al. 2016), o OrthoMaM oferece alinhamentos de sequências nucleares identificadas como ortólogas entre si, dado um grupo de táxons referência escolhido pelo usuário entre 43 gêneros de Mammalia. Como visto anteriormente, sequências ortólogas são as ideais para inferência de árvores filogenéticas, pois evitam conflito topológico provocado por duplicação gênica seguida do uso inadvertido de cópias parálogas (Degnan e Rosenberg 2009).

Dos alinhamentos disponíveis no banco de dados, foram baixados apenas aqueles que contivessem sequências ortólogas em relação a membros representantes de cada uma das 14 ordens de placentários e 3 ordens de marsupiais disponíveis no banco de dados. Entre os 43 gêneros disponíveis, cada gênero escolhido como representante foi, até a versão 9 do OrthoMaM, o único membro disponível para sua ordem ou aquele com maior número de ortólogos em relação aos demais membros. Foram eles *Monodelphis, Sarcophilus, Macropus, Echinops, Loxodonta, Procavia, Choloepus, Dasypus, Erinaceus, Pteropus, Equus, Canis, Bos, Oryctolagus, Mus, Tupaia* e *Homo*. Um total de 3.398 alinhamentos de sequências atenderam aos critérios de ortologia em relação a esses 17 gêneros.

Para evitar artefatos topológicos resultantes da atração de ramos longos entre o grupo externo e táxons com curto tempo de geração no grupo interno (fenômeno observado em análises preliminares), eliminei as sequências de *Ornitorhinchus* - o único membro da classe Monotremata no OrthoMaM - de todos os alinhamentos que as continham. Os alinhamentos resultantes representaram entre 33 e 42 gêneros de Mammalia (mediana e moda 41), dos quais 3 gêneros de Marsupialia assumiram a função de grupo externo.

A referência taxonômica utilizada para delimitar as ordens de Mammalia foi a terceira edição do livro *Mammal Species of the World* (Wilson & Reeder 2005), com apenas duas exceções. Na primeira, considerei os gêneros *Erinaceus* e *Sorex* como membros da mesma ordem (Eulipotyphla), contrastando com a distinção entre ordens Erinaceomorpha e Soricomorpha feita neste livro. Por sua vez, enquanto o *Mammal Species of the World* considera os grupos Artiodactyla e Cetacea como ordens separadas, aqui também designei seus membros como parte da mesma ordem (Cetartiodactyla). Essas resoluções seguem um acúmulo de evidências moleculares para o parafiletismo de Soricomorpha e Artiodactyla, respectivamente, que vem ocorrendo desde o final dos anos 90, a partir dos trabalhos de Douady e colaboradores (2002) e Gatesy e colaboradores (1999).

3.3.2 – Processamento das sequências e escolha do modelo de substituição

Todas as sequências disponibilizadas para download pelo OrthoMaM passam por um processamento prévio, sendo possível obtê-las apenas alinhadas ou já processadas para remoção de sítios pouco informativos e sequências espúrias, selecionados automaticamente. Para maior controle dos parâmetros envolvidos na amostragem dos dados, obtive os alinhamentos dos 3.398 marcadores selecionados sem qualquer processamento prévio.

O algoritmo do OrthoMaM compõe os alinhamentos através do programa MAFFT (Katoh e Standley, 2013) e os refina alinhando as sequências por seus códons no MACSE (Ranwez et al. 2011), evitando, assim, o efeito de eventuais *frameshifs* e erros de sequenciamento ou de montagem (Douzery et al. 2014). Em uma fase adicional de processamento dos alinhamentos, eliminei as regiões flanqueantes de cada alinhamento (sítios com mais de 25% de *missing data*), além de sítios internos que contivessem menos de três caracteres que não fossem *gaps*, mantendo, dessa forma, somente aqueles minimamente informativos.

Como visava utilizar esses dados empíricos para avaliar o performance do modelo de regressão em testes topológicos e, por sua vez, o modelo de regressão foi estimado de dados simulados sob a matriz de substituição GTR, selecionei para demais análises apenas os marcadores empíricos para os quais o GTR tinha melhor adequação. Para identificar esses marcadores, cada alinhamento processado passou por um teste de modelo de substituição no programa IQ-tree, no qual o critério de decisão utilizado foi o AICc (*corrected Akaike Information Criterion* – Hurvich e Tsai 1993). A preferência pelo AICc se deu pelo fato de apresentar menor probabilidade de seleção de modelos excessivamente complexos em relação ao AIC simples e, ao mesmo tempo, por não pressupor que o modelo de substituição correto

esteja entre os modelos candidatos, como faz o BIC (*Bayesian Information Criterion*) (Burnham e Anderson 2004).

Nesta etapa, contemplando todas as opções de modelos disponíveis para teste no IQ-tree, o AICc selecionou entre diferentes combinações de 22 matrizes de substituição e 3 níveis de representação da heterogeneidade nas taxas de substituição ao longo dos sítios: nenhuma variação (i.e. taxa constante para todos os sítios); quatro categorias de variação gama (4Γ) ou quatro categorias gama mais uma categoria de sítios invariáveis (4Γ+I). Todos os 1343 alinhamentos para os quais GTR pontuou mais alto nesse critério seguiram para inferência de suas árvores de gene.

3.3.3 – Inferências filogenéticas e testes topológicos

Toda a etapa de inferência das árvores de gene se deu também por critério de ML no IQ-tree e sob a matriz de substituição GTR, com número de categorias gama correspondente à indicação do teste de modelo para cada caso. Naqueles em que a indicação foi GTR ou GTR+4Γ, a inferência se deu sob um modelo de mesma estrutura; já nos que foi GTR+4Γ+I, a árvore foi inferida utilizando GTR+8Γ, pelos mesmos motivos de economia computacional apontada anteriormente: trocando uma categoria invariável por mais categorias gama, dispensa-se a estimação da proporção de sítios invariáveis e do alfa simultaneamente.

Para testar uma árvore qualquer contra resoluções alternativas das relações entre seus grupos de táxons, é preciso que o monofiletismo de cada um desses grupos, bem como do grupo externo, sejam pressupostos razoáveis. Visando garantir essa condição para os testes da raiz de Placentalia, identifiquei e triei, das árvores de gene inferidas, aquelas que recuperaram as quatro superordens de Placentalia e o grupo externo Marsupialia como grupos monofiléticos. Dos 3.398 genes iniciais, apenas 787 mostraram ter, individualmente, sinal suficiente para recuperar o monofiletismo desses grupos; prosseguindo para uma nova leva de inferências.

Nessa segunda etapa, entretanto, estabeleci *constraints* nas topologias iniciais (i.e. não permiti que fossem otimizadas) em 9 pontos específicos: nos 2 nós referentes ao MRCA do grupo externo e ao do grupo interno; nos 4 nós referentes ao MRCA de cada uma das quatro superordens do grupo interno e nos 3 nós que determinam a ordem das especiações entre as linhagens das superordens e, por conseguinte, a posição da raiz de Placentalia. Como para a última existem 15 resoluções possíveis (incluindo as três mais aceitas, vistas anteriormente), realizei 15 corridas independentes sob diferentes *constraints,* para cada um dos 787 genes. Otimizadas todas as árvores alternativas e resolvidos, assim, o restante de suas topologias e

seus comprimentos de ramos; aquela que pontuou maior verossimilhança foi comparada às demais via testes topológicos realizados, também, para cada gene independentemente.

Do processamento dos alinhamentos do OrthoMaM, passando pela sua triagem através da seleção de modelos e inferência das árvores de gene, até a composição dos *constraints* e otimização por máxima verossimilhança, todas as etapas foram automatizadas através de mais um *pipeline* em Shell (**Suplementar 3**) que agrega scripts em Perl, em R e linhas de comando do IQ-tree. Em seguida, foram executados os testes topológicos de Kishino e Hasegawa (KH), Shimodaira e Hasegawa (SH), o *Approximatelly Unbiased* (AU) no IQ-tree e o *Equidistant Delta* (ED) através de uma função customizada em R.

Para a obtenção dos p-valores em todos os testes topológicos, exceto o ED, foram geradas 10.000 réplicas via RELL. Já a aplicação do ED seguiu a concepção apresentada na **seção 3.2**, com previsão de ICs através do modelo regredido. Em todos os testes, o nível de confiança fixado para rejeição ou não das hipóteses nulas foi de 95%. Os testes via KH e ED não corrigem para múltiplas comparações; sendo realizados, nesses casos, par a par, entre a árvore de maior verossimilhança e cada uma das demais em cada conjunto de árvores testadas.

O IQ-tree calcula as verossimilhanças das árvores em cada conjunto testado com base nos parâmetros de substituição otimizados com apenas uma delas (inferida por ML, por máxima parcimônia ou fixada *a priori*), não sendo possível especificar o conjunto de parâmetros otimizados anteriormente com cada árvore. Portanto, a cada grupo de árvores, foi necessário fixar como referência aquela que pontuou maior verossimilhança antes dos testes. A verossimilhança das árvores e seus p-valores podem ter sido afetados por essa aproximação, havendo até, em alguns casos, mudança da topologia que obteve maior verossimilhança (i.e. alteração da árvore de gene).

Além das análises de árvores de genes, realizei inferências de um concatenado dos mesmos 787 *loci* finais. Esse concatenado, totalizando pouco mais de 2,6 milhões de bases, foi montado automaticamente através de um script em *Perl* e analisado no IQtree para inferência da árvore filogenética em três diferentes níveis de complexidade: análise não-particionada, análise particionada simples (permite-se diferentes taxas de substituição, mas tamanhos de ramos proporcionais entre partições) e análise particionada sensível a heterotaquia (permite, adicionalmente, que os tamanhos de ramos variem de forma independente para cada partição).

O intuito com essas três análises foi avaliar, além do efeito do concatenamento sobre a resolução da raiz, o da complexidade do modelo evolutivo utilizado; uma vez demonstrado

que todos os *loci* incluídos possuem sinal suficiente para o monofiletismo das superordens. Para fins de redução de custo computacional, o esquema de particionamento utilizado nos dois últimos casos foi selecionado previamente pelo IQtree entre os 10% melhores esquemas (opção *-rcluster 10*) de combinação entre os diferentes *loci*. A inclusão, no concatenado, de apenas genes para os quais GTR demonstrou o melhor ajuste anteriormente também permitiu acelerar a etapa de seleção de modelo (entre GTR+4Γ ou GTR+8Γ) para cada partição. Na análise não particionada, o modelo utilizado foi GTR+12Γ, para melhor contemplar a heterogeneidade entre os milhões de sítios do concatenado.

Demais avaliações dos resultados, tanto das árvores dos genes quanto do concatenado, foram realizadas no R, contando principalmente com testes de Kolmogorov-Smirnov para comparar distribuições e verificar a relação entre a composição dos alinhamentos e os resultados obtidos para a raiz de Placentalia. O objetivo primário com esses procedimentos foi investigar a presença de artefatos topológicos supostamente gerados por vieses composicionais; efeito reportado anteriormente por Romiguier e colaboradores (2013).

4 – Resultados

4.1 – Da influência das variáveis observadas sobre a distribuição de Δln*L*

4.1.1 – Análises de dispersão

As análises iniciais dos valores de Δln*L*, entre árvores equidistantes das que geraram os dados, demonstraram que tendem a se dispersar em torno de 0 nessa condição. Entre a proporção de *gaps*, extensão do alinhamento, número de ramos na árvore, forma da distribuição gama (representado pelo alfa estimado) ou distâncias BSD - tanto entre árvores alternativas quanto entre elas e a árvore verdadeira correspondente -, nenhuma das variáveis exerceu qualquer influência sobre a média da distribuição de Δln*L* (**Figura 19**); sempre que respeitada uma diferença máxima de distância de 0.01 BSD entre cada árvore comparada (alternativas) e a verdadeira. Por si só, esse resultado representou um prognóstico positivo para a viabilidade do teste topológico idealizado, já que a condição de equidistância se refletiu em equidade de adequação das árvores comparadas **-** quando medida pelas diferenças entre suas verossimilhanças para um grande número de observações (alinhamentos).

Por outro lado, a tendência de difusão (afastamento) dos valores Δln*L* de 0 pareceu ser afetada por algumas dessas variáveis. Entre elas, a proporção de gaps (**Figura 19a**) e a extensão do alinhamento (**Figura 19b**) mostraram clara relação negativa com a difusão; sendo menos linear para o número de sítios. Apesar de ocorrer ligeira alteração na difusão ao longo de diferentes números de ramos nas árvores comparadas (**Figura 19c**) e valores de alfa (**Figura 19d**), não se viu, *a priori*, relação monotônica (consistentemente positiva ou negativa) com estas variáveis. Depois, uma análise dos histogramas de todas as variáveis mostrou que a discreta variação com alfa estava associada, na verdade, a uma melhor representação de alfas entre 0.6 e 0.75, mas não explicou a errática variação de Δln*L* com diferentes números de ramos. Por sua vez, tanto as distâncias (BSD) entre as árvores comparadas (**Figura 19e**) quanto a média de suas distâncias para sua árvore verdadeira (**Figura 19f**) mostraram relação muito similar e positiva com a difusão dos Δln*L*s.

O que torna este último resultado interessante, por destacar a relevância das distâncias topológicas para comparar a adequação de árvores conflitantes, é também o que o tornou preocupante para a concepção do modelo de regressão idealizado inicialmente. Como visto na seção **3.2**, a proposta do modelo era prever o desvio padrão da distribuição de Δln*L* na condição de igual adequação (equidistância) das árvores comparadas, dispensando qualquer

informação *a posteriori*, fora as variáveis preditoras no modelo e o próprio Δln*L* (a estatística teste). Ao lidar com dados empíricos, no entanto, cuja árvore verdadeira é desconhecida, não se pode contar com a distância entre ela e as árvores comparadas como uma dessas variáveis; apenas com a distância *entre* as árvores comparadas. Buscando solucionar esse paradoxo, avaliei os limites da influência de uma sobre a outra.

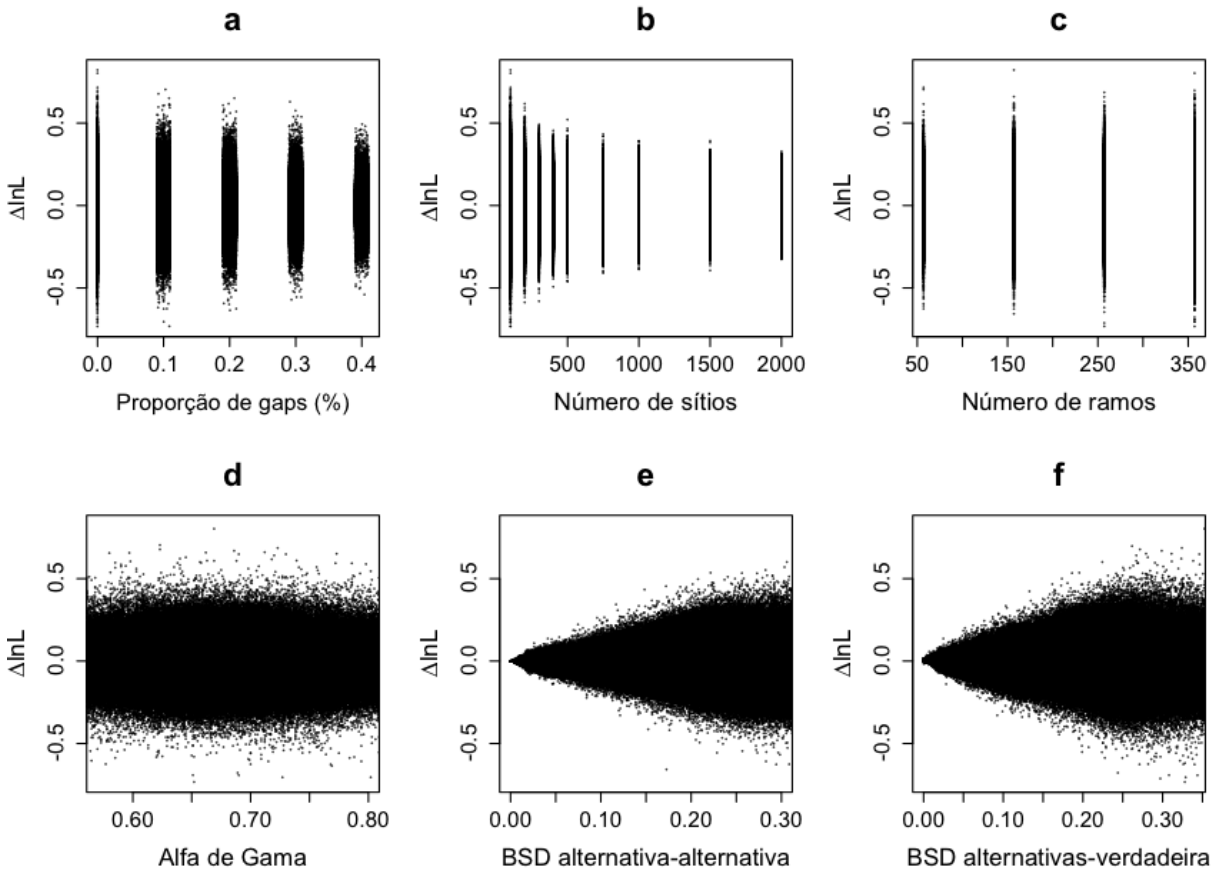


Figura 19: Dispersão dos valores de Δln*L* (médias dos valores de Δln*L* por sítio) entre árvores alternativas que se mantiveram equidistantes da árvore verdadeira correspondente. Os plots mostram os efeitos de diferentes variáveis sobre a distribuição de Δln*L* nessa condição. Entre elas, estão os fatores que foram totalmente controlados ao longo das simulações de dados e otimizações das árvores (proporção de gaps (a), número de sítios no alinhamento (b) e o número de ramos nas árvores - totalmente resolvidas - que foram comparadas (c)), como também a distância entre essas árvores (e), parcialmente controlada - antes da otimização de seus ramos. Adicionalmente, os efeitos de variáveis não controladas, como o parâmetro alfa (d), que determina a forma da distribuição gama, e distância média entre as árvores alternativas e sua árvore verdadeira (f), também foram representados. Para todas as variáveis contínuas (d, e, f), as observações apresentadas são de seções que correspondem ao HPD 90% (highest probability density) aproximado empiricamente para suas distribuições; o restante das observações foi omitido para evitar a percepção artefatual de alteração da dispersão Δln*L*, causada por subrepresentação.

Um *plot* das distâncias entre árvores alternativas contra suas distâncias médias para as árvores verdadeiras indicou que, em meu conjunto de dados, há certa tendência de correlação positiva entre elas, mas com múltiplos casos de violações dessa orientação (**Figura 20a**). Como esperado, a coerência geométrica da BSD impede que a primeira distância seja maior que o dobro da segunda; da mesma forma que seria impossível a base de um triângulo ser maior que a soma dos seus dois outros lados. Entretanto, assim como a soma desses outros lados poderia se estender indefinidamente (não importando o tamanho da base), o mesmo ocorre para a distância da árvore verdadeira, em alguns casos variando para muito além da distância entre alternativas. Visto isso, a única forma de executar a correta regressão do modelo sem incluir essa variável foi reduzir a amostra de 2.989.720 observações para as 1.539.734 cujas árvores alternativas apresentaram distâncias em relação à verdadeira iguais ou menores que a distância em relação a seu par (**Figura 20b**).

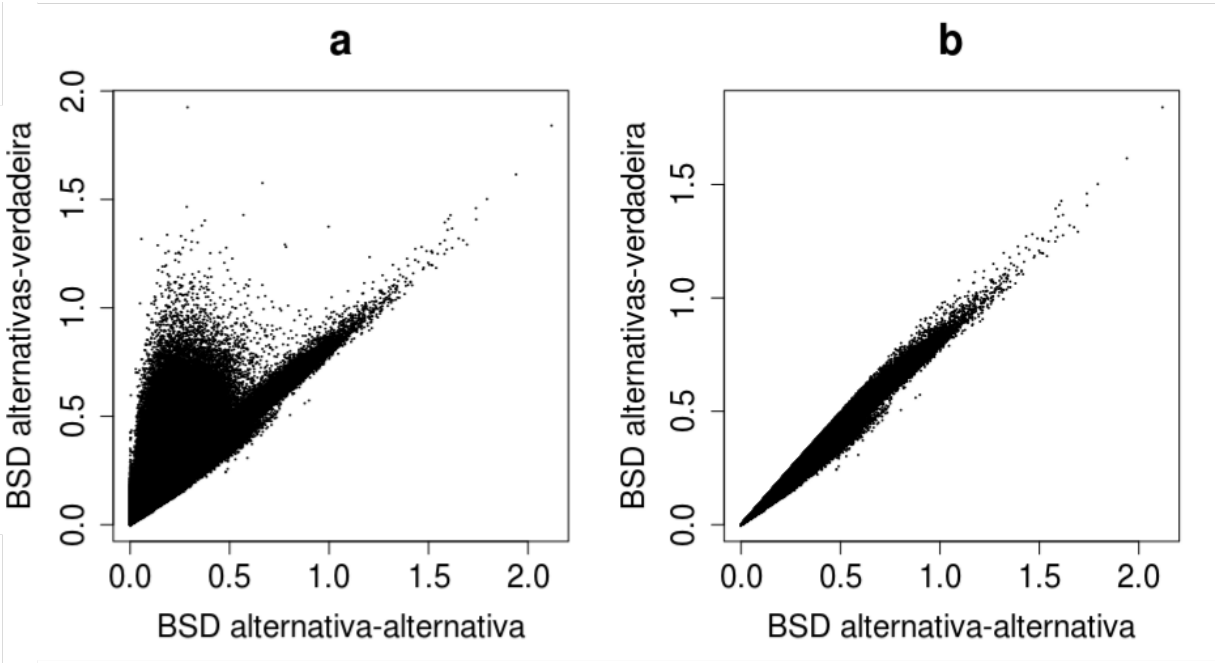


Figura 20: Relação da distância BSD entre árvores alternativas com a média de suas distâncias para a verdadeira correspondente, a) para todos os casos nos quais os pares de árvores alternativas se mantiveram equidistantes da verdadeira após otimização dos comprimentos de ramos e b) após redução desse conjunto de observações àquelas árvores cujas distâncias da verdadeira não ultrapassaram as distâncias em relação a seus pares.

Todavia, essa redução da amostra exigiu a inclusão de mais um pressuposto à aplicação do modelo: testa-se a equidistância das árvores comparadas em relação à verdadeira desde que essa distância não supere a distância entre as árvores testadas. Apesar disso potencialmente lesar a confiabilidade do teste para casos gerais, ainda pode reservar sua

utilidade na investigação de pontos de conflito topológico no limiar entre filogenias próximas, como é o caso do problema da raiz de Placentalia. Ainda, com a limitação das distâncias em relação à árvore verdadeira, tornou-se mais clara a existência de uma relação positiva entre o número de ramos nas árvores e a difusão dos valores de Δln*L* (**Figura 21**); antes mascarada por padrões distintos de variação dessa distância entre topologias de diferentes tamanhos.

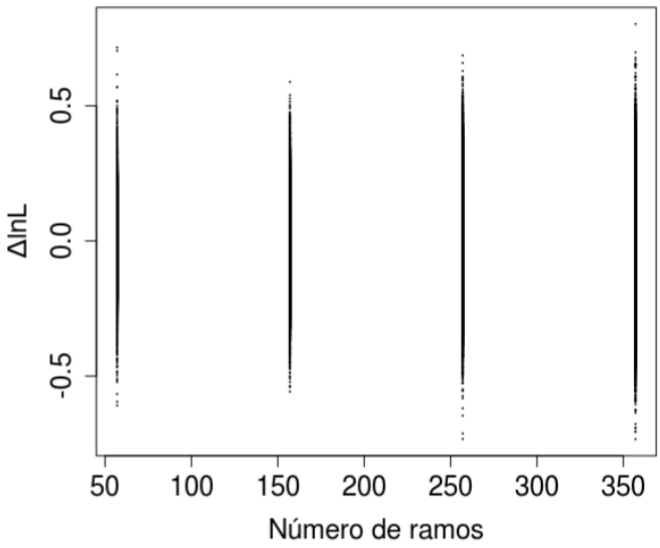


Figura 21: Dispersão de Δln*L* com diferentes números de ramos, após redução da amostra. Torna-se um pouco mais clara a existência de uma tênue relação positiva entre o número de ramos e a difusão dos valores de Δln*L* em relação a 0.

4.1.2 – Análises de regressão e o modelo preditor

Após a obtenção de 1.080 distribuições com 200 valores de Δln*L* (uma para cada combinação possível entre diferentes categorias das quatro variáveis controladas), foi possível atestar de que forma cada variável afeta o desvio padrão dessas distribuições; no caso, a estatística que o modelo de regressão se propõe a prever. Os boxplots resultantes endossaram a maior parte dos resultados das análises de dispersão de Δln*L*. Vê-se que tanto a distância entre as árvores comparadas quanto seu número de ramos têm relação aproximadamente linear e positiva com o desvio padrão (SD) de Δln*L* (**Figura 22a e d**). Já o número de sítios e a proporção de gaps no alinhamento tem relação negativa com o SD (**Figura 22b e c**); sendo a primeira a única que apresenta padrão não-linear, tendendo a ser menos expressiva à media que aumenta a extensão do alinhamento (**Figura 22b**).

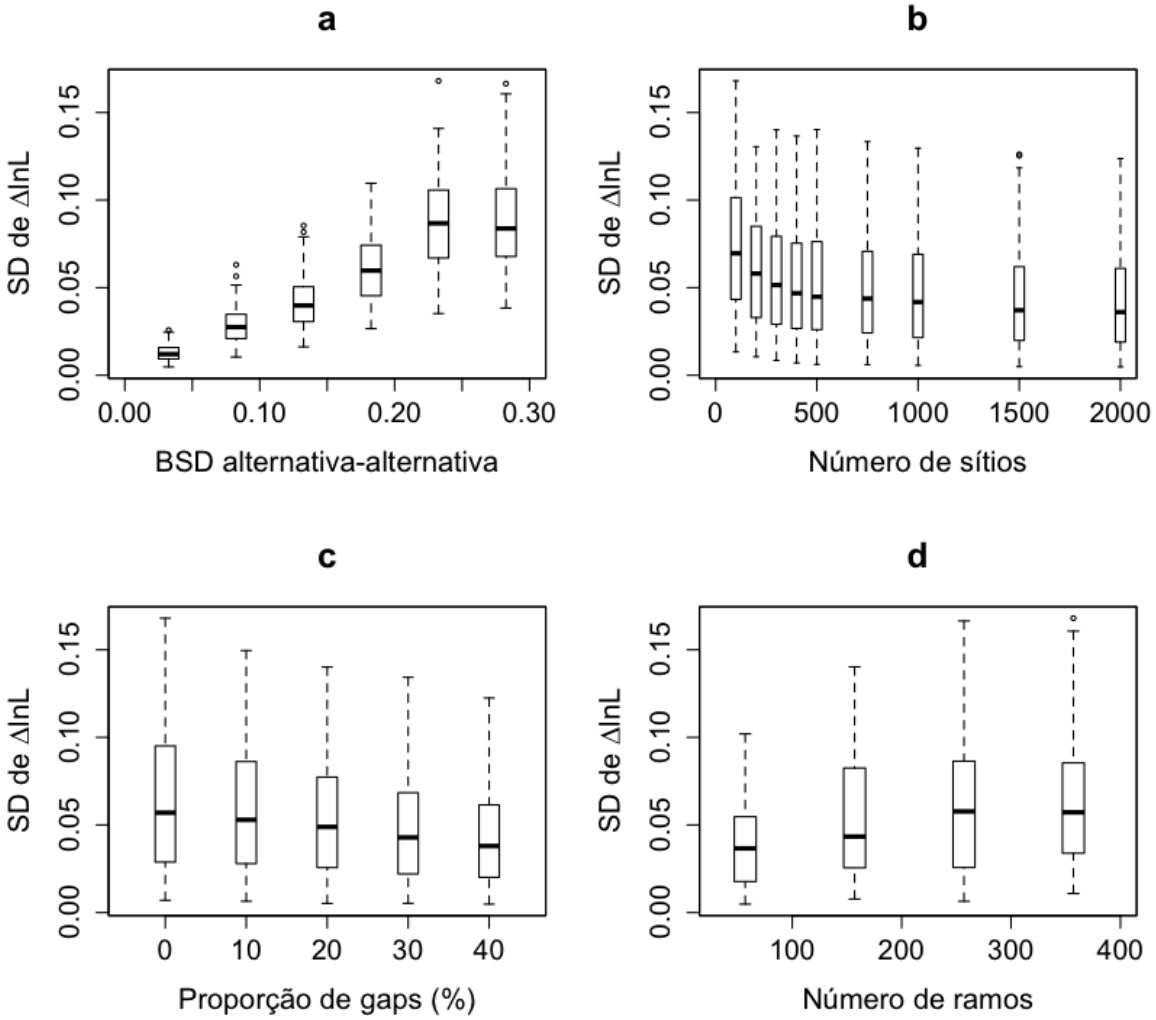


Figura 22: Relação entre as variáveis parcial ou totalmente controladas e o desvio padrão (SD) de Δln*L*. Cada boxplot representa a distribuição de SDs das distribuições de Δln*L* que se encaixam numa determinada categoria de a) distâncias BSD entre as árvores comparadas, b) número de sítios, c) proporção de gaps nos alinhamentos ou d) número de ramos na árvore. Uma vez que o número de categorias difere entre variáveis, o número de observações representadas em cada boxplot também varia, mas é equivalente entre boxplots ao longo da mesma variável.

Entre as variáveis que demonstraram padrão linear com o SD, análises preliminares de regressão simples confirmaram correlações significativas. A distância entre as árvores comparadas demonstrou ter a maior influência sobre o desvio padrão de Δln*L* (R^2 = 0,68) entre as analisadas, seguida do número de ramos na árvore (R^2 = 0,067) e da proporção de gaps (R^2 = 0,049). No esforço para adequar todas as variáveis aos pressupostos da regressão múltipla, a logaritmização natural, tanto dos valores dessas variáveis (independentes) quanto do desvio padrão das distribuições de Δln*L* (variável dependente), foi a melhor forma de promover uma relação linear entre a dependente e cada uma das independentes (**Figura 23**)

simultaneamente. A única exceção ocorreu com a proporção de *gaps*, que já mostrou relação linear com o logaritmo do desvio padrão sem necessitar qualquer transformação (**Figura23c**).

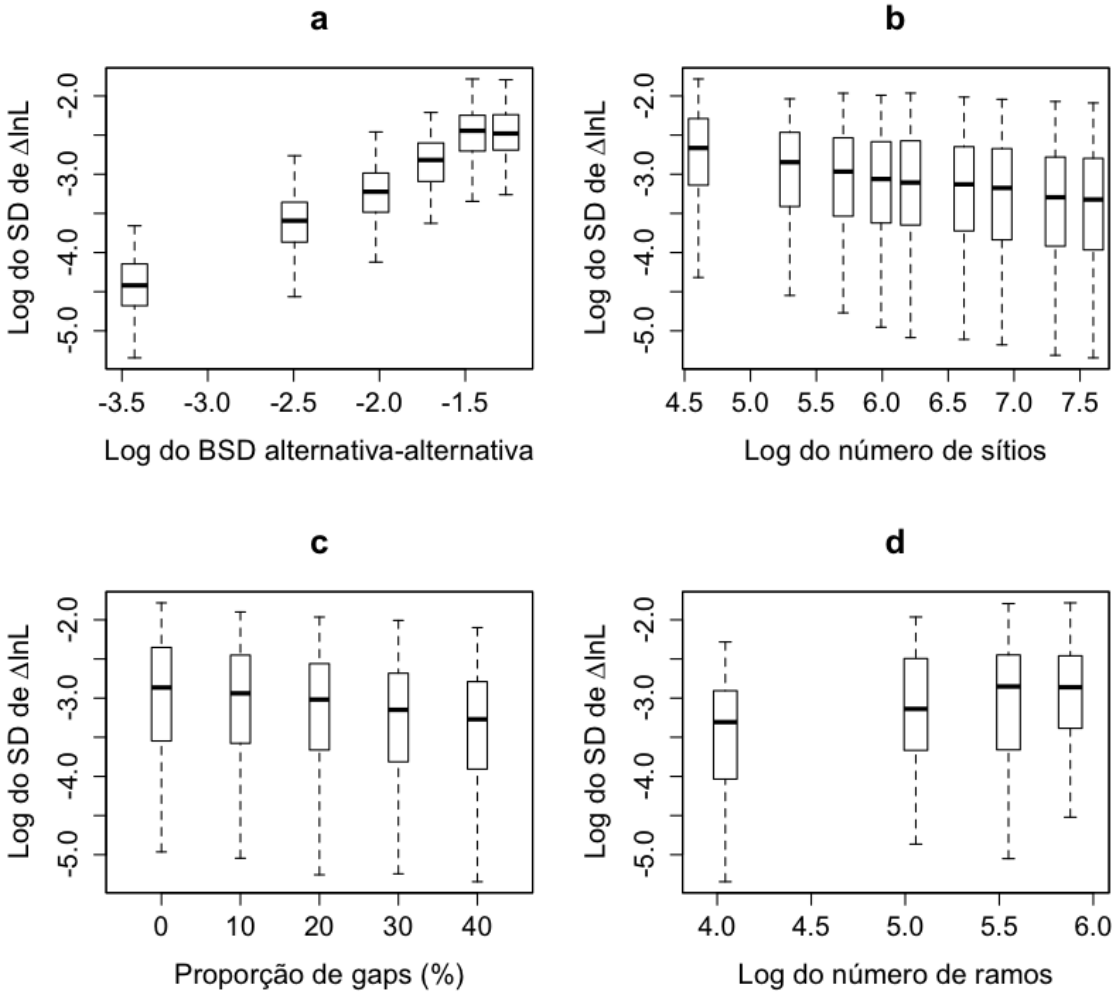


Figura 23: Relação entre as variáveis parcial ou totalmente controladas e o desvio padrão (SD) de Δln*L* após transformações finais. Equivale à figura 22, mas após logaritmização de todas as variáveis, exceto da proporção de gaps.

Da variável que apresentou anteriormente o maior valor de R^2 via regressão simples até a que obteve o menor, um processo de inclusão sucessiva de cada uma das variáveis transformadas ao modelo de regressão levou à obtenção de valores progressivamente maiores de R^2 pelo modelo de regressão múltipla. O logaritmo da distância entre árvores alternativas foi o que explicou a maior parte da variação no logaritmo do desvio padrão, atingindo R^2 de 0.79. Com adição do logaritmo do número de ramos ao modelo, houve melhora do R^2 para 0.86 e assim sucessivamente; até a adição da proporção de *gaps*, que representou o menor aprimoramento no modelo, de 0.919 para 0.954 no valor do R^2.

As etapas de adição de variáveis ao modelo foram realizadas em concomitância com análises da homocedasticidade e da normalidade de seus resíduos. O modelo mais complexo, que incluiu os quatro fatores parcial ou totalmente controlados foi o que, além de atingir R^2 mais alto (0.954), possibilitou a maior aproximação do padrão homocedástico (**Figura 24a**) e da normalidade dos resíduos (**Figura 24b e c**). O erro padrão total dos resíduos desse modelo foi 0.1648, sobre 1.075 graus de liberdade (1080 observações menos 5 parâmetros - um intercepto e quatro coeficientes de regressão). Todos os coeficientes de regressão entre as variáveis independentes e a dependente foram significativos pelo teste *t* ($p < 2e\text{-}16$). Por esses motivos, esse foi o modelo selecionado para implementação no novo teste topológico.

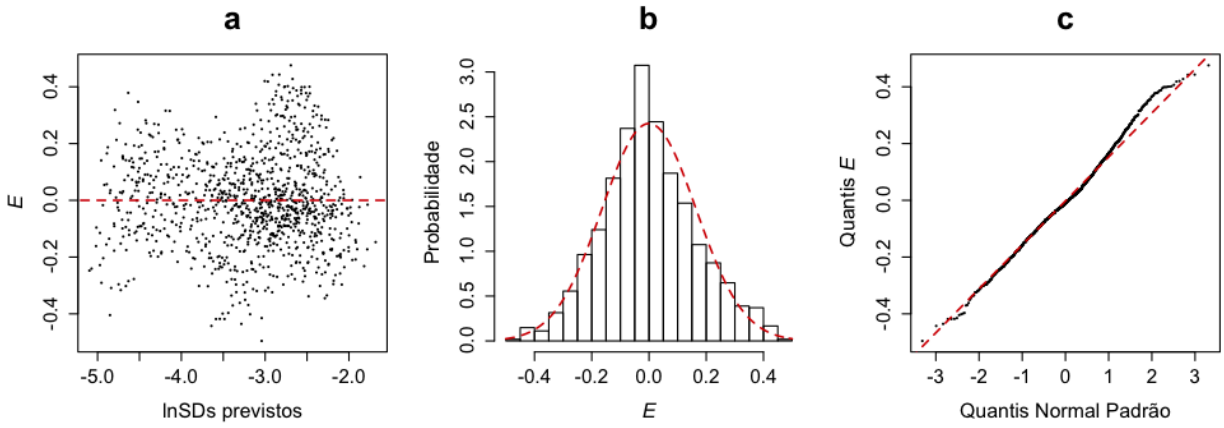


Figura 24: Adequação dos resíduos (E) do modelo selecionado aos pressupostos da regressão múltipla: Homocedasticidade, representada pela (a) dispersão dos resíduos para os diferentes lnSDs (logaritmos do desvio padrão de $\Delta\ln L$) previstos pelo modelo, em relação ao resíduo 0 (tracejado horizontal em vermelho) e normalidade multivariada, verificável pelo (b) histograma da distribuição de resíduos com sobreposição da curva de densidade de probabilidade normal de média e desvio padrão correspondentes (tracejada em vermelho) ou pelo (c) plot dos quantis "teóricos" - da normal padrão - contra os quantis empíricos da distribuição de resíduos, quando comparado à tendência esperada com uma normal de média e desvio padrão correspondentes (diagonal tracejada em vermelho).

Aplicado ao teste *Equidistant Delta*, o modelo de regressão múltipla assumiu a fórmula

$$lnSD = 0.945 \times lnBSD_{t1t2} + 0.295 \times lnK_{t1t2} - 0.201 \times lnS_m - 1.009 \times GP_m - 1.292$$

em que *lnSD* é o logaritmo natural do desvio padrão de $\Delta\ln L$, *lnBSD* é o logaritmo natural da distância topológica (BSD) entre as árvores comparadas, *lnK* é o logaritmo natural do número de parâmetros (ramos) nas árvores - que deve ser o mesmo em ambas - , *lnS* é o logaritmo do

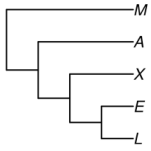
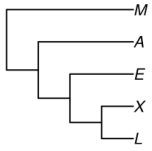
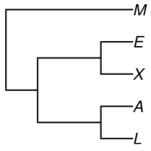
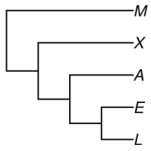
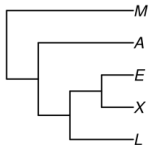
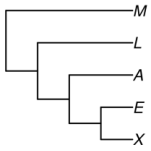
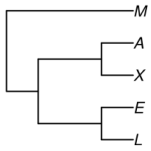
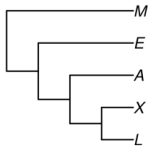
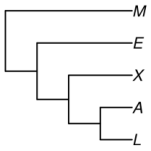
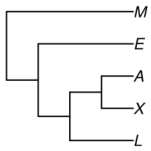
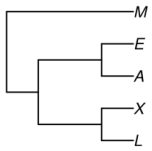
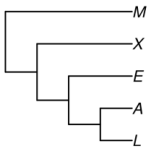
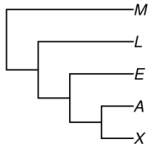
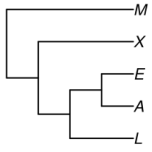
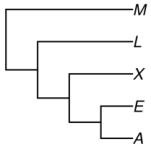
número de sítios e *GP* é a proporção de *gaps* no alinhamento. No caso, *t1* e *t2* é o par de árvores comparadas no teste e *m* é o alinhamento utilizado para otimizar seus comprimentos de ramos, demais parâmetros e calcular suas verossimilhanças. Elevando o número neperiano (2.718282) ao lnSD previsto pelo modelo, obtive o SD de $\Delta \ln L$ esperado na condição de equidistância para cada caso testado mais tarde.

4.2 – Das resoluções para a raiz de Placentalia pelos genes amostrados

4.2.1 – Árvores de genes

Das 787 árvores de genes inferidas dos *loci* selecionados, aproximadamente 96% recuperaram uma das três resoluções mais aceitas para a posição da raiz de Placentalia, com Boreoeutheria como grupo monofilético. Nessa parcela, as três ocorreram em frequências similares; a maioria das árvores de gene (270) posicionando Afrotheria como grupo irmão do restante da diversidade na subclasse, seguida de 254 que colocaram Xenarthra nesta posição. Em menor frequência, mas não menos expressiva, 231 árvores de gene uniram Afrotheria e Xenarthra no clado Atlantogenata.

Os pouco mais de 4% restantes não recuperaram o monofiletismo de Boreoeutheria e variaram entre sete diferentes resoluções. Destes, 25 *loci* (ou 3% do total) recuperaram Atlantogenata, posicionando ora Laurasiatheria, ora Euarchontoglires na posição de superordem mais externa em Placentalia. Por fim, aproximadamente 1% dos genes (7 árvores) não reconstruíram Boreoeutheria nem Atlantogenata. Versões colapsadas de cada uma das resoluções possíveis entre as quatro superordens, bem como suas frequências entre as árvores de genes podem ser verificadas no **Quadro 1**, junto às notações pelas quais passam a ser referidas doravante.

Notação	Topologia	N	Notação	Topologia	N	Notação	Topologia	N
T1		270	**T6**		3	**T11**		0
T2		254	**T7**		1	**T12**		0
T3		231	**T8**		1	**T13**		0
T4		13	**T9**		1	**T14**		0
T5		12	**T10**		1	**T15**		0

Quadro 1: **Todas as relações possíveis entre as superordens de Placentalia, em ordem de ocorrência entre as árvores de gene recuperadas, via máxima verossimilhança.** A notação indicada para cada topologia é aquela que será utilizada para referir ao grupo de árvores que a recuperaram, enquanto o N indica o tamanho desses grupos entre os 787 loci analisados. Nos terminais de cada topologia representada, M remete ao grupo externo Marsupialia; A à superordem Afrotheria; X a Xenarthra; E a Euarchontoglires e L a Laurasiatheria.

4.2.2 – Busca por artefatos

Em 2013, Romiguier e colaboradores apontaram alguns fatores que estariam associados à inferência artefatual de Boreoeutheria + Atlantogenata (T3), em detrimento de Boreoeutheria com Afrotheria como superordem mais externa (T1). Os autores argumentam que *loci* ricos em GC nas terceiras posições de códon são sintomáticos de *hot-spots* de recombinação e que elevadas taxas de substituição nessas regiões aumentariam sua tendência a saturação; ou seja, a falha do modelo de substituição em inferir as múltiplas substituição ocorridas. Entre as consequências dessa característica para a árvore de Placentalia, destacam o prolongamento dos ramos que levam a Afrotheria e Xenarthra, gerando o efeito de "atração de ramos longos"; isto é, sua união em um grupo monofilético baseada em homoplasias. Além

disso, supõem a possibilidade de que maiores concentrações GC estejam associadas a redução do tamanho dos genes, junto de seu poder de resolução filogenética.

Para determinar se a ocorrência de alguma das resoluções mais frequentes nas árvores de genes obtidas aqui poderia ser artefato de um desses vieses, procurei por características dos genes analisados - e dos parâmetros evolutivos estimados através deles - que pudessem estar mais frequentemente associadas a T1, a T2 ou a T3. A começar pela proporção GC nas terceiras posições de códon (GC3), análises de histogramas para os conjuntos de *loci* que recuperaram as três resoluções não indicaram, a princípio, uma tendência expressiva de viés topológico causado por essa variável (**Figura 25**).

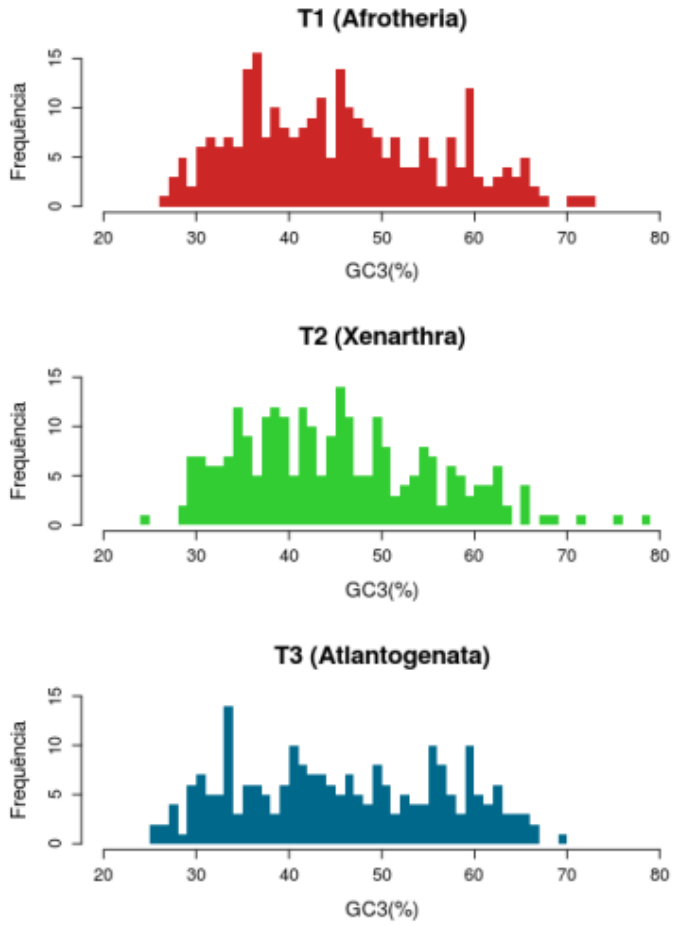


Figura 25: Proporções GC nas terceiras posições de códons (GC3) dos loci que recuperaram cada uma das três topologias mais frequentes. Em vermelho, genes cujas árvores tiveram raiz entre Afrotheria e o restante da diversidade de Placentalia (T1); em verde, entre Xenarthra e os demais (T2); por fim, em azul, entre Atlantogenata e Boreoeutheria (T3).

A fim de determinar se as distribuições de GC3 diferiram significativamente para cada posição recuperada para a raiz, empreguei testes bicaudais de Kolmogorov-Smirnov (KS). Sendo um teste não-paramétrico, que compara pares de distribuições empíricas pela diferença

entre suas funções de distribuição acumulada (FDAs), o KS não indicou diferença significativa entre nenhum dos três pares possíveis (com nível de confiança 95%); mesmo sem qualquer correção para múltiplas as comparações. Assumindo, como hipótese nula, que foram amostradas de uma mesma população, o teste entre as GC3 de T1 e de T2 obteve p-valor de 0,965; entre as distribuições para T1 e para T3, p-valor igual a 0,332; já entre T2 e T3, p-valor 0,209.

Ainda, para testar a alegação de que marcadores ricos em AT tem maior tendência a recuperar Boreoeutheria e Afrotheria externo e, os ricos em GC, um vies para Boreoeutheria com Atlantogenata (Romiguier et al. 2013), executei, também, um KS monocaudal, tendo como hipótese alternativa que a FDA para T1 está significativamente acima da FDA para T3 (i.e. T1 apresentam proporções GC3 significativamente menores que T3). O resultado deste teste também indicou diferença não-significativa ($p = 0,167$).

Visto que não há relação entre a proporção GC3 e viés topológico em meu conjunto de dados e considerada a possibilidade da relação detectada por Romiguier e coautores (2013) ter sido artefato de uma correlação não observada entre o número de sítios e a proporção GC3 em seus marcadores; investiguei também se o tamanho dos *loci* se associava de alguma forma às topologias recuperadas aqui. Uma análise de histogramas dos tamanhos dos genes analisados indicou, novamente, diferenças pouco expressivas entre os que recuperaram as topologias T1, T2 e T3 (**Figura 26**). Apesar do grupo de genes de T1 apresentar extensões extremas (entre 10 e 20 mil *bp*) pouco mais frequentemente que os demais, essa observação poderia ser também efeito da maior frequência de árvores de gene T1 em relação às demais.

Para definir se os marcadores T1 seriam de fato significativamente mais longos que os de T2 ou T3, apliquei os mesmos testes de KS, dessa vez comparando o grupo T1 a cada um dos demais via testes monocaudais. Ao testar a hipótese nula de equidade das distribuições contra a hipótese alternativa da FDA de T1 se posicionar abaixo da de T2 (i.e. genes significativamente mais longos em T1), a diferença se mostrou não-significativa ($p = 0,881$). Da mesma forma, o teste entre a FDA de T1 e de T3 não rejeitou a hipótese de equidade de suas distribuições ($p = 0,904$).

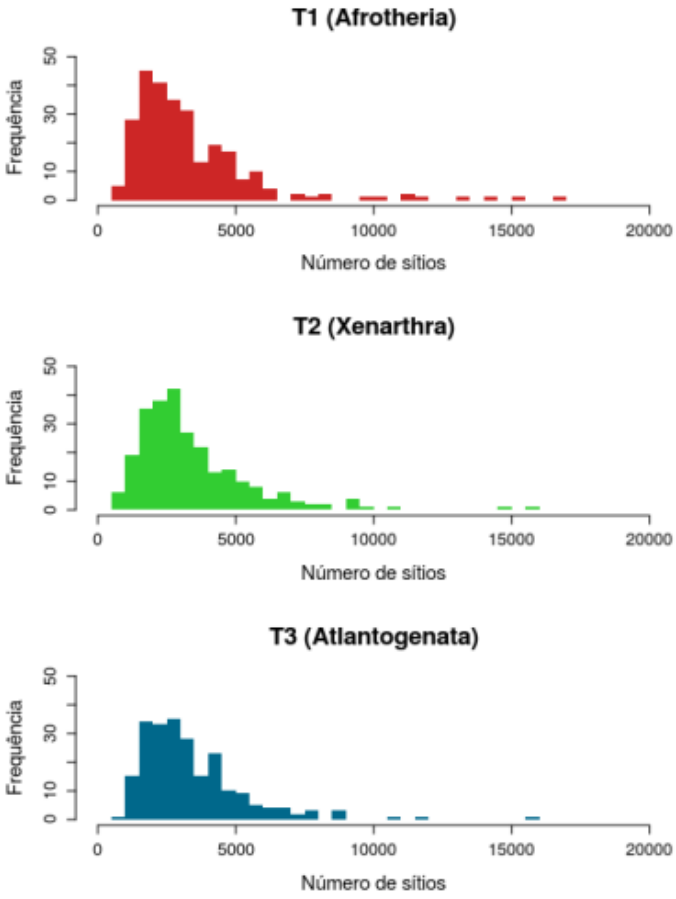


Figura 26: Número de sítios nos loci que recuperaram cada uma das três principais topologias. As cores seguem o padrão da figura 25: em vermelho, genes cujas árvores tiveram raiz entre Afrotheria e as demais superordens (T1); em verde, entre Xenarthra e o restante (T2); em azul, entre Atlantogenata e Boreoeutheria (T3).

Mesmo que a proporção GC3 e o número de sítios nos *loci* analisados não estejam associados à recuperação de uma ou outra topologia em especial, a união das superordens Afrotheria e Xenarthra em Atlantogenata ainda poderia resultar de fenômenos de atração de ramos longos em minhas inferências. Se este for o caso, espera-se que as distâncias filogenéticas (i.e. acúmulo de substituições por sítio), entre o ancestral comum mais recente (MRCA) de Placentalia e os MRCAs de Afrotheria e Xenarthra, sejam relativamente altos se comparados às distâncias para os MRCAs das demais superordens.

Para verificar essa possibilidade, tratei distâncias filogenéticas em termos da somas dos comprimentos de ramos do nó ancestral imediato de Placentalia até o nó imediato de cada superordem, em cada árvore. Com esses valores, obtive a razão da distância média para Xenarthra e Afrotheria ((Plac.-Xen. + Plac.-Afr.) / 2) pela distância média até todas as superdordens ((Plac.-Xen. + Plac.-Afr. + Plac.-Laur. + Plac.-Euarc.) / 4). Quanto maior essa

razão (que chamo aqui de ARS ou Acúmulo Relativo de Substituições) para uma dada árvore, maior seria seu acúmulo relativo de substituições para Afrotheria e Xenarthra. Avaliando as distribuições dos ARSs obtidos para as árvores de genes, verifica-se que os das árvores T1 e T2 mantiveram média próxima a 1, enquanto as T3 demonstraram viés para valores superiores (**Figura 27**)

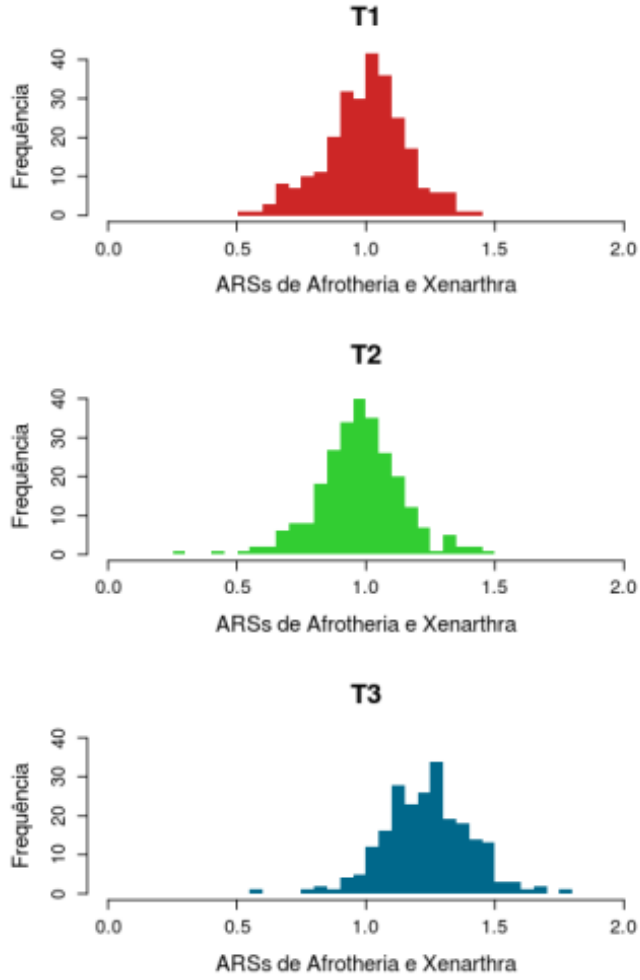


Figura 27*:* ARSs de Afrotheria e Xenarthra nas árvores que obtêm cada uma das três resoluções mais frequentes. Cada ARS equivale à razão da distância filogenética média entre o MRCA de Placentalia e os MRCAs de Xenarthra e de Afrotheria pela distâncias média da mesma distância para o MRCA de cada uma das quatro superordens. Em outras palavras, é um indicativo do Acúmulo Relativo de Substituições até Xenarthra e Afrotheria em cada árvore; quanto maior o ARS, mais relativamente alto é esse acúmulo. Em vermelho, distribuição para as árvores que tiveram raiz entre Afrotheria e o restante da diversidade de Placentalia (T1); em verde, entre Xenarthra e os demais (T2); em azul, entre Atlantogenata e Boreoeutheria (T3).

Novos testes KS monocaudais, entre a distribuição de ARS para T3 e as demais, indicaram que seu FDA fica significativamente abaixo (ou seja, tem ARSs superiores), tanto do FDA de T1 ($p = 2.2e-16$), quanto de T2 ($p = 2.2e-16$). Mesmo uma correção de Bonferroni

para as múltiplas comparações mantém esses p-valores muito abaixo do nível de significância corrigido (0,025). Este resultado reforça que, nos genes que recuperam Atlantogenata, ocorreu acúmulo relativamente alto de substituições ao longo das linhagens que levam a Xenarthra e Afrotheria; causa em potencial de atração de ramos longos, caso os modelos de substituição utilizados para esses genes não tenham sido sofisticados o bastante para evitar saturação.

4.2.3 – Análises concatenadas

Como visto na introdução, análises concatenadas tendem a negligenciar a hetererogeneidade de sinais entre os múltiplos marcadores analisados. Por outro lado, contam com quantidades maiores de observações; sendo a princípio capazes de comportar uma maior parametrização do modelo evolutivo sem grandes detrimentos para a precisão pelo acúmulo de erro randômico na estimação de cada parâmetro. Considerando isso, realizei três inferências filogenéticas a partir do concatenado dos 787 genes, com três diferentes níveis de parametrização, no intento de observar o efeito de modelos progressivamente complexos sobre a recuperação da raiz de Placentalia.

Surpreendentemente, apesar de ter sido a resolução menos frequente entre os três principais posicionamentos da raiz nas árvores de gene, a análise concatenada e sem qualquer particionamento recuperou Atlantogenata e Boreoeutheria (**Figura 28**). No entanto, como os resultados anteriores demonstraram, há condições no conjunto de dados para que essa resolução tenha sido efeito de atração de ramos longos entre Xenarthra e Afrotheria (T3). Apesar de esperar-se que as 12 categorias gama utilizadas nessa análise acomodem razoavelmente bem os sítios de evolução mais rápida, por si só podem ter falhado em contornar a saturação em alguns deles.

Porém, modelos mais sofisticados, como os particionados, poderiam minorar a influência de *loci* de evolução mais rápida sobre a inferência, ao permitirem maior liberdade de variação das taxas de substituição nesses grupos de genes. Não obstante, a análise particionada e com comprimentos de ramos proporcionais entre partições recuperou exatamente a mesma topologia que a não particionada; não só em relação à posição da raiz de Placentalia, mas a todos os demais parentescos na árvore. Não bastando, nem mesmo a análise mais paramétrica realizada - particionada com comprimentos de ramos independentes entre partições - discordou da anterior em qualquer ponto.

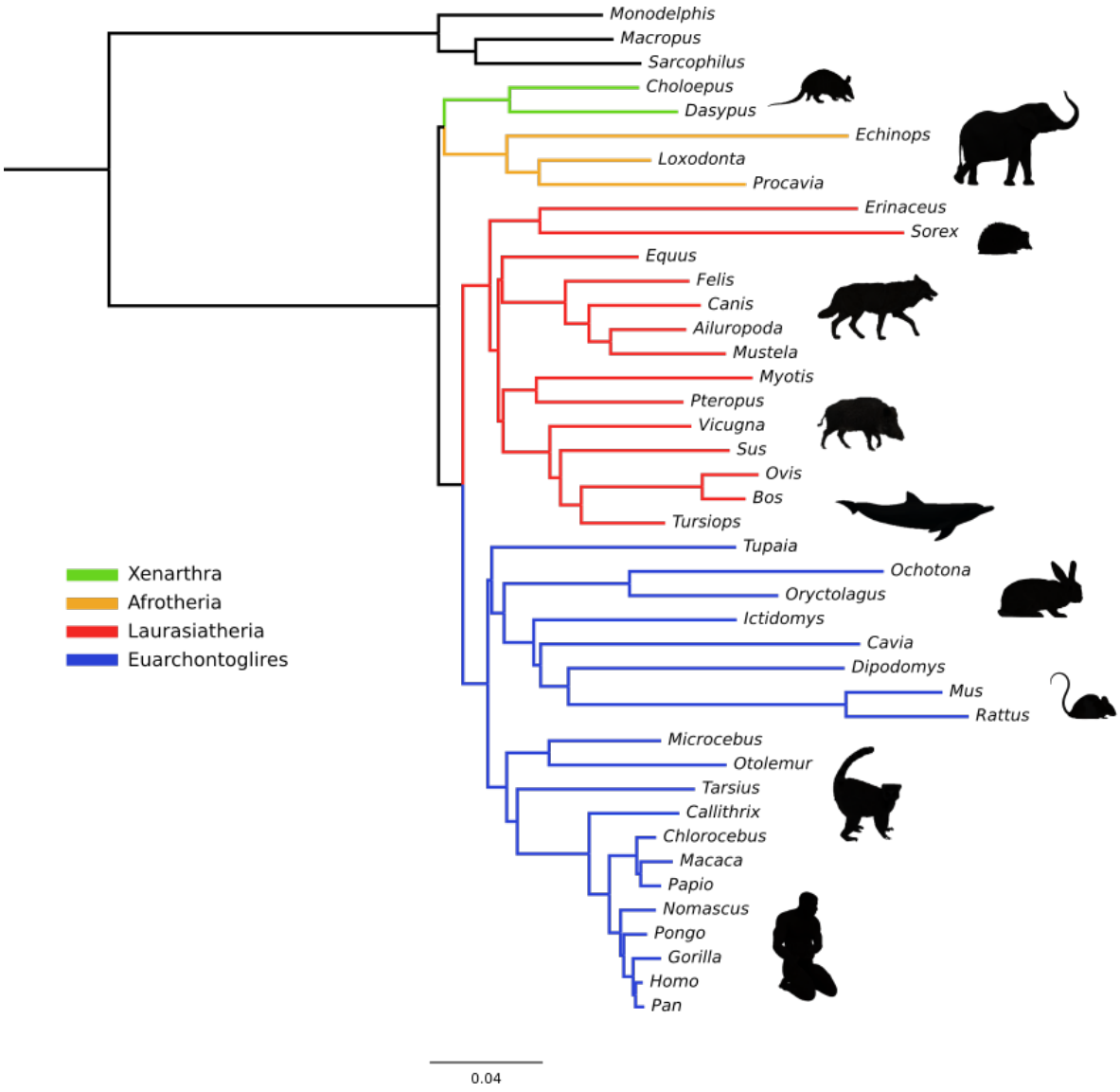


Figura 28: Árvore de máxima verossimilhança para 42 gêneros de Mammalia com base no concatenado de 787 genes nucleares (2.627.166 bp). Resultou de uma análise não-particionada, assumindo uma única matriz de substituição GTR (melhor ajuste para todos os genes incluídos) e taxas de substituição em distribuição gama com 12 categorias de variação.

4.2.4 – Testes topológicos

A similaridade das frequências de T1, T2 e T3 entre as árvores de gene inferidas e a aparente inocuidade do particionamento para as inferências concatenadas nada mais do que reiteram a dificuldade de resolução da raiz de Placentalia, mas são resultados limitados pela confiança em estimativas pontuais. Os testes topológicos, em contrapartida, permitiram pôr essa confiança a prova ao determinar a significância do sinal contido em cada gene para sua resolução de ML. Analisando os grupos de *loci* T1, T2 e T2 pelas suas proporções de rejeição de cada uma das 14 demais resoluções possíveis para as relações entre superordens, foi possível verificar variações na força decisória do sinal contido em cada um.

Para os testes ED e KH monocaudais (i.e. realizados entre árvores de ML e cada uma das demais) ficou evidente uma acentuada dificuldade de rejeição de suas hipóteses nulas (de equidistância em relação à verdadeira e de esperança nula para a diferença de verossimilhança, respectivamente), quando a comparação se deu com qualquer uma das outras resoluções principais (T1, T2 ou T3) (**Tabela 1**). Nesses casos, as proporções de p-valores abaixo de 0,05 variaram de 0 a um máximo de 2,6% nos testes ED e de 1,1% até 3,1% via KH. Em ambos os tipos de testes, genes T1 pareceram rejeitar a hipótese nula em prol dessa estimativa com frequência um pouco menor que os demais grupos (médias aproximadamente 6% abaixo de T2 em ED e de T3 em KH).

Tabela 1 Proporções (%) de rejeição da hipótese nula nos testes monocaudais ED e KH. Para cada árvore de ML nos grupos T1, T2 e T3 (colunas), realizou-se um teste ED e um teste KH comparando-a par a par com cada uma das árvores restritas a resoluções diferentes das relações entre superordens (fileiras). O valor indicado em cada célula da tabela é a porcentagem dos p-valores resultantes que ficou abaixo de 0,05. Para o ED, o valor representa o número de vezes, a cada 100 genes, em que a árvore ML (T1, T2 ou T3) e seu par não são equidistantes da árvore verdadeira; favorecendo a hipótese alternativa de que a ML esteja mais próxima. Segundo o KH, indica a porcentagem de vezes em que a verossimilhança da ML foi significativamente maior que a de seu par. Nas representações em newick de cada árvore alternativa, *L* refere a Laurasiatheria, *E* a Euarchontoglires, *X* a Xenarthra e *A* a Afrotehria.

	ED			KH		
	T1	T2	T3	T1	T2	T3
T1: (((L,E),X),A)	-	1.2	2.6	-	2.7	2.2
T2: (((L,E),A),X)	0.0	-	1.3	1.5	-	1.3
T3: ((L,E),(X,A))	0.0	0.0	-	1.1	3.1	-
T4: ((L,(X,A)),E)	15.9	19.3	10.0	52.2	58.8	38.3
T5: (((X,A),E),L)	15.6	20.1	9.5	53.7	60.3	37.9
T6: (((L,X),E),A)	69.3	70.1	80.5	62.2	80.2	81.9
T7: ((L,(X,E)),A)	68.9	70.9	81.0	63.3	79.8	81.9
T8: (((L,X),A),E)	63.3	72.8	83.5	72.2	82.5	81.9
T9: ((L,X),(A,E))	64.1	74.4	84.0	74.8	80.9	81.9
T10: ((L,(A,E)),X)	64.8	79.1	83.1	71.9	72.0	83.3
T11: ((L,A),(X,E))	64.4	72.8	82.7	73.0	80.2	82.4
T12: (((X,E),A),L)	64.8	74.0	84.0	72.2	79.4	83.3
T13: (((L,A),X),E)	65.2	72.8	81.0	73.3	82.1	81.1
T14: (((L,A),E),X)	64.1	79.1	82.7	74.4	71.6	80.2
T15: (((A,E),X),L)	67.0	72.4	84.8	74.1	78.6	82.8
Média	49.1	55.6	60.8	58.6	65.2	64.3

Ainda, nota-se que comparações com T4 e T5 (as únicas resoluções além de T3 que recuperam Atlantogenata) rejeitam a hipótese nula menos frequentemente que comparações com as demais árvores que quebram o monofiletismo de Boreoeutheria (T6 até T15). Apesar de evidente nos testes KH, esse fenômeno é bem mais expressivo nos resultados do ED, apresentando proporções de rejeição menores que a metade de suas médias (**Tabela 1**). Considerando o caráter experimental do ED e o fato de que os resultados do KH monocaudal não são decisivos em caso de rejeição da hipótese nula - dado seu viés de seleção - foi necessário considerar, também, os resultados de testes que comprovadamente controlam com sucesso suas taxas de erro do tipo I, como o SH e o AU.

Mais uma vez, viu-se baixíssimas frequências de rejeição entre as árvores T1, T2 e T3 pelo SH e pelo AU (**Tabela 2**); mostrando que, entre os genes analisados, há pouquíssimos capazes de reiterar suas estimativas de ML diante das outras hipóteses principais para a raiz de Placentalia. Adicionalmente, vê-se que os genes dos grupos T1 e T2 demonstraram, também, relativa dificuldade de rejeição das árvores que, apesar de não recuperarem o monofiletismo de Boreoeutheria, mantiveram Afrotheria (T6 e T7) ou Xenarthra (T10 e T14), respectivamente, na posição de grupo irmão do restante dos placentários. Essa tendência é observável apenas com os testes KH, SH e AU, parecendo se inverter quando aplicado o ED. No entanto, efeito análogo ocorre para os genes T3 em todos os testes, rejeitando bem menos frequentemente as árvores que recuperam Atlantogenata (T4 e T5), independente de suas resoluções para as demais relações.

Esta última tendência é particularmente interessante, porque mais uma vez, com os testes SH e AU, não se limitou aos genes T3 (Boreoeutheria + Atlantogenata), ocorrendo também com aqueles que tem T1 e T2 como estimativas de ML (**Tabela 2**). Além disso, as frequências de rejeição de T4 e T5 não só estão abaixo das médias de rejeição em todos os testes e por todos os três grupos de genes, como também estão abaixo das proporções de rejeição de T6 e T7 por genes T1 e de T10 e T14 por genes T2; o que pode indicar a existência de um "sinal críptico" para Atlantogenata em alguns *loci* que não o recuperaram em suas estimativas pontuais.

Tabela 2: Proporções (%) de rejeição da hipótese nula nos testes SH e AU. Para cada árvore de ML nos grupos T1, T2 e T3 (colunas), foi feito um teste SH e um AU comparando-a a todas as árvores forçadas a resoluções alternativas (fileiras). Novamente, o valor indicado em cada célula da tabela é a porcentagem dos p-valores resultantes que ficou abaixo de 0.05. Segundo o SH, representa a porcentagem de vezes em que as árvores de resolução igual à da fileira correspondente não representaram a evolução do gene tão bem quanto as demais. Já pelo AU, das vezes em que o valor esperado para sua verossimilhança (considerada a curvatura de suas fronteiras topológicas) não foi maior nem igual aos esperados para as demais árvores.

	SH			AU		
	T1	T2	T3	T1	T2	T3
T1: (((L,E),X),A)	-	0.0	0.9	-	6.2	5.7
T2: (((L,E),A),X)	0.0	-	0.9	5.6	-	5.7
T3: ((L,E),(X,A))	0.0	0.0	-	4.4	5.1	-
T4: ((L,(X,A)),E)	25.6	29.6	6.6	60.7	65.8	44.1
T5: (((X,A),E),L)	25.9	30.0	6.6	60.4	65.4	44.1
T6: (((L,X),E),A)	51.1	73.9	73.6	68.1	86.4	85.0
T7: ((L,(X,E)),A)	51.1	72.4	74.9	68.5	85.6	84.6
T8: (((L,X),A),E)	66.3	76.7	75.8	83.7	88.3	88.1
T9: ((L,X),(A,E))	67.4	75.5	76.7	83.0	88.7	87.2
T10: ((L,(A,E)),X)	67.8	56.8	73.1	81.9	75.9	89.4
T11: ((L,A),(X,E))	66.3	76.3	76.2	86.3	86.8	86.8
T12: (((X,E),A),L)	66.3	75.1	78.0	84.1	85.6	89.0
T13: (((L,A),X),E)	67.4	75.1	73.6	87.4	85.2	87.2
T14: (((L,A),E),X)	67.4	59.5	73.6	81.9	76.3	85.5
T15: (((A,E),X),L)	67.4	73.2	77.1	83.7	83.3	89.0
Média	49.3	55.3	54.8	67.1	70.3	69.4

Levando em conta que os resultados vistos até agora podem estar sujeitos ao nível de significância escolhido (0,05) e visando investigar em maior profundidade a hipótese do sinal críptico para Atlantogenata, avaliei as distribuições de p-valores obtidos, dos grupos T1 e T2, para as árvores alternativas via teste AU - já conhecido por seu bom controle das taxas de erro, tanto do tipo I quanto II. Tanto avaliações de histogramas, quanto da obliquidade (*skewness*) dessas distribuições e testes de Kolmogorov-Smirnov foram empregados para compará-las.

No que se referia aos genes que recuperam T1 (Boreoeutheria e Afrotheria externo), foquei especialmente nas diferenças entre suas distribuições de p-valor para as árvores T4 e T5 (Atlantogenata, sem Boreoeutheria) e para T6 e T7 (Afrotheria externo, sem Boreoeutheria), para determinar se os p-valores para as primeiras foram de fato significativamente menores que para as últimas. Pela análise de histogramas, todos os grupos

demonstraram distribuições similares (**Figura 29**), mas com valores de obliquidade pouco maiores para T6 e T7 (Afrotheria externo) em comparação a T4 e T5 (Atlantogenata). Quatro testes de KS entre esses dois pares de distribuições indicaram que T4 e T5 obtiveram p-valores significativamente maiores dos genes que recuperam T1 como ML (p-valor total (KS x 4) = 0,009).

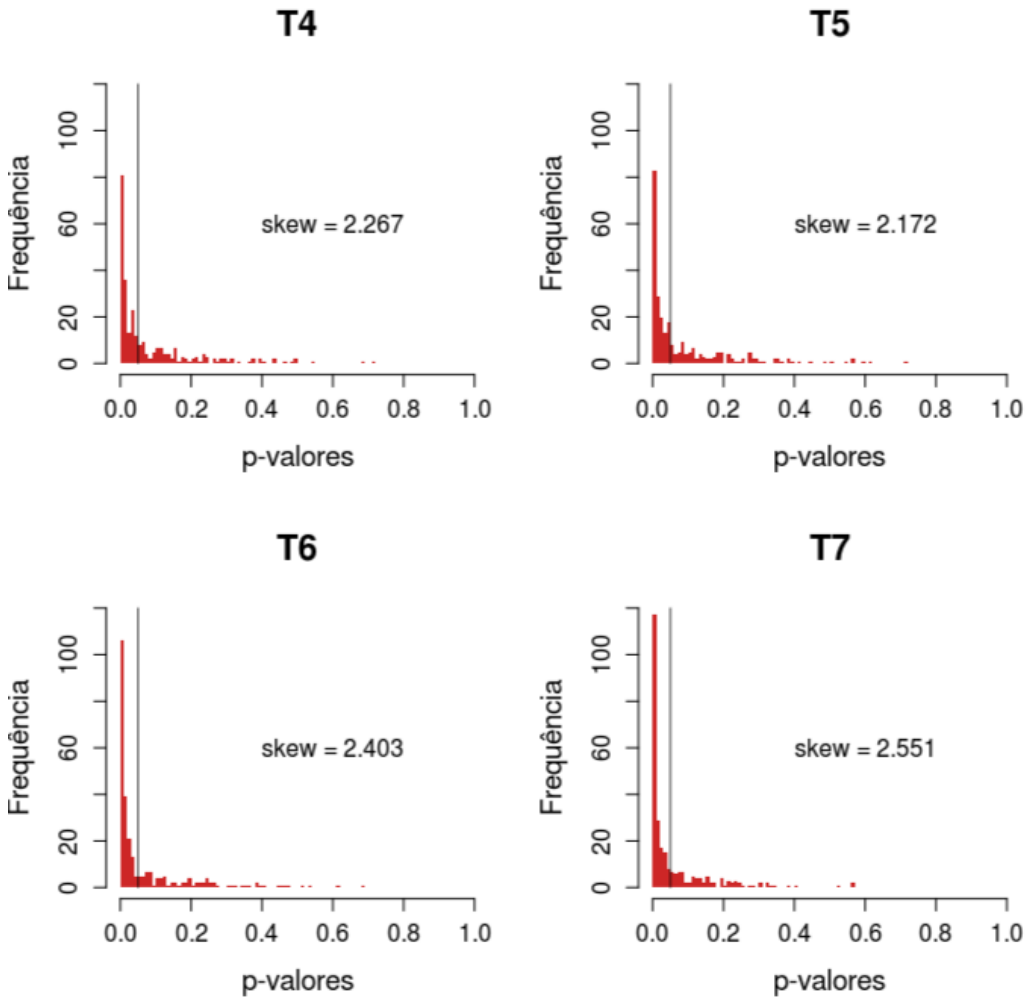


Figura 29: Distribuição dos p-valores (via AU) para as árvores T4, T5, T6 e T7, pelos genes que recuperam T1 (Boreoeutheria monofilético e Afrotheria como superordem mais externa) como árvore ML. T4 e T5 fixam Atlantogenata, mas rompem Boreoeutheria posicionando suas superordens em diferentes posições; enquanto T6 e T7 rompem ambos, mas fixam Afrotheria como grupo irmão do restante da diversidade placentária. A linha vertical negra marca o nível de significância assumido anteriormente (0,05). Junto a cada histograma está indicada também a obliquidade (*skewness*) de cada distribuição; quanto maior a obliquidade, maior sua assimetria, concentrando-se à esquerda (p-valores menores).

Já com os p-valores dos genes que reconstroem T2 (Boreoeutheria e Xenarthra externo), avaliei suas distribuições para T4 e T5, mas, dessa vez, comparando-as às de T10 e T14 - as demais árvores, além de T2, que recuperam Xenarthra como superordem mais externa, apesar de quebrarem o monofiletismo de Boreoeutheria. Mais uma vez, as distribuições de T3 e T4 obtiveram valores de obliquidade relativamente inferiores (**Figura 30**), porém com diferenças menores para os de T10 e T14 que as observadas anteriormente. No entanto, novos testes de KS confirmaram, com ainda mais significância, que os p-valores de T4 e T5 foram maiores que os de T10 e T14, dados os genes que recuperam T2 como estimativa pontual (p-valor total (KS × 4) = 0,0001).

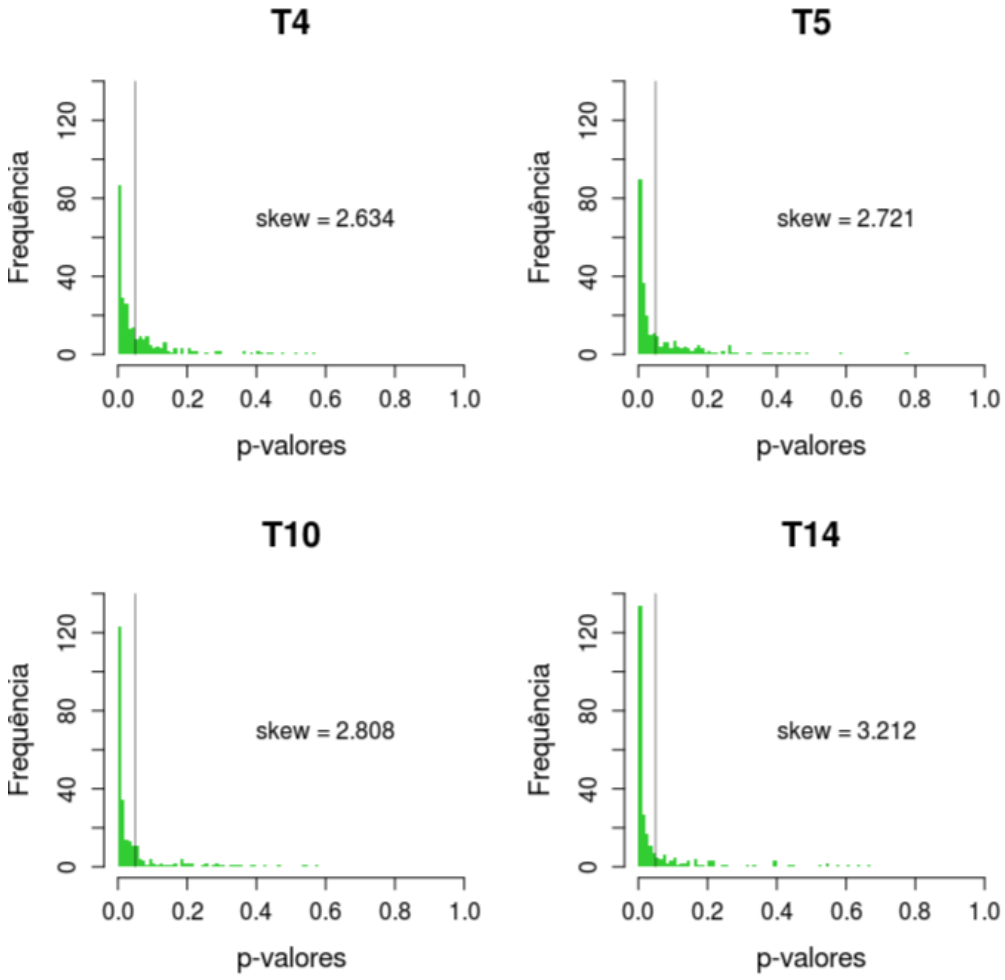


Figura 30: Distribuição dos p-valores (via AU) para as árvores T4, T5, T10 e T14, obtidos pelos genes que recuperam T2 (Boreoeutheria monofilético e Xenarthra como superordem mais externa) como árvore ML. T4 e T5 fixam Atlantogenata, mas rompem Boreoeutheria posicionando suas superordens em diferentes posições; já T10 e T11 rompem ambos, mas fixam Xenarthra na posição de grupo irmão das demais superordens. Novamente, a linha vertical negra indica o nível de significância considerado nos testes anteriormente (0,05) e, o valor de 'skew', a obliquidade de cada distribuição. Quanto maior a obliquidade, maior foi a assimetria da distribuição, tendendo a concentrar-se à esquerda (p-valores menores).

4.3 – Da performance do *Equidistant Delta*

Como visto pelas proporções médias de rejeição da hipótese nula (**Tabelas 1 e 2**), no geral o ED aparentou relativo conservadorismo, rejeitando com frequências menores que o KH ou o AU, mas não tanto a ponto de apresentar médias menores que o SH; teste que tem, sabidamente, um viés conservador (taxas altas de erro do tipo II), apesar de controlar suas taxas de erro do tipo I. Apesar disso, o ED concordou com a tendência de todos os outros testes em quase nunca rejeitar a H0 quando confrontado com alguma das três resoluções principais (T1, T2 e T3), bem como em rejeitar Atlantogenata (em T3, T4 e T5) com frequências relativamente menores que resoluções adversas.

Por outro lado, o ED foi na contramão dos três testes tradicionais quando rejeitou as árvores T5 e T6, contra as árvores ML T1, e também T7 e T14, contra as T2, mais frequentemente que o restante das árvores que não fixaram Boreoeutheria (**Tabela 1**). Tentando compreender esse fenômeno, reavaliei todas as proporções de rejeição (via ED) já vistas, junto às medianas das distâncias topológicas (BSD) entre as árvores de ML e as resoluções alternativas (**Tabela 3**). Dado que as distâncias topológicas foram as variáveis que melhor explicaram a variação na distribuição de $\Delta \ln L$ (**Figura 19**), o aumento das proporções de reijeição dessas árvores pelo teste ED poderia estar associado a distâncias BSD menores entre elas e as árvores ML. Na prática, entretanto, apesar de maior similaridade entre as árvores comparadas ter implicado maior dificuldade de rejeição da hipótese nula em alguns casos (e.g. T1 vs T3 ou T3 vs. T5), essa relação de fato se inverte quando na comparação de T1 com T6/T7 ou T2 com T10/T12 (**Tabela 3**).

Como para cada rodada de 14 testes (entre a ML e as demais resoluções) foi analisado um mesmo *locus,* de tamanho e proporção de *gaps* fixos e, também, porque todas as árvores apresentaram o mesmo número de ramos, não há motivos para suspeitar que as demais variáveis contempladas pelo modelo do ED estejam associadas à "inversão" vista. Ainda, dado que os testes clássicos não indicaram diferenças de verossimilhança especialmente maiores entre árvores T1 e T6/T7 ou entre T2 e T10/T14 (**Tabelas 1 e 2**), é possível que o fracasso do ED em apresentar proporções de rejeição condizentes, nesses casos, esteja associado à negligência da distância entre essas árvores e a árvore verdadeira, desconhecida.

Tabela 3: Comparativo entre as proporções de rejeição (%) pelo teste ED e as medianas das distâncias topológicas (BSD), entre as árvores que obtiveram maior verossimilhança mais frequentemente entre os genes analisados (colunas) e as que foram limitadas a cada uma das resoluções alternativas (fileiras).

	T1		T2		T3	
	ED	BSD	ED	BSD	ED	BSD
T1: (((L,E),X),A)	-	-	1.2	0.008	2.6	0.006
T2: (((L,E),A),X)	0.0	0.007	-	-	1.3	0.006
T3: ((L,E),(X,A))	0.0	0.007	0.0	0.008	-	-
T4: ((L,(X,A)),E)	15.9	0.017	19.3	0.018	10.0	0.012
T5: (((X,A),E),L)	15.6	0.017	20.1	0.018	9.5	0.012
T6: (((L,X),E),A)	69.3	0.013	70.1	0.018	80.5	0.015
T7: ((L,(X,E)),A)	68.9	0.013	70.9	0.018	81.0	0.015
T8: (((L,X),A),E)	63.3	0.019	72.8	0.019	83.5	0.016
T9: ((L,X),(A,E))	64.1	0.019	74.4	0.019	84.0	0.016
T10: ((L,(A,E)),X)	64.8	0.017	79.1	0.012	83.1	0.015
T11: ((L,A),(X,E))	64.4	0.019	72.8	0.019	82.7	0.016
T12: (((X,E),A),L)	64.8	0.018	74.0	0.019	84.0	0.016
T13: (((L,A),X),E)	65.2	0.019	72.8	0.019	81.0	0.016
T14: (((L,A),E),X)	64.1	0.017	79.1	0.013	82.7	0.015
T15: (((A,E),X),L)	67.0	0.019	72.4	0.019	84.8	0.016

Apesar de suas possíveis falhas, o teste ED demonstrou a eficiência computacional desejada ao dispensar a necessidade de replicações dos valores de $\Delta \ln L$. Os 11.018 testes ED executados (entre cada uma das 787 árvores de maior verossimilhança e das 14 demais) foram completados em 146 segundos, utilizando um único núcleo de um processador Intel® Core™ i7 (2,50GHz). Em contraste, os testes KH, SH e AU com as mesmas árvores, que foram executados simultaneamente no IQ-tree, dividindo as mesmas réplicas de RELL e contando com paralelização em dois núcleos do mesmo processador, foram concluídas em aproximadamente 2.500 segundos (pouco mais de 40 minutos).

5 – Discussão

5.1 – Previsibilidade da distribuição de Δln*L* e a viabilidade do ED

Como comentado anteriormente, a maioria dos testes topológicos desenvolvidos até hoje depende, em alguma etapa, do procedimento de *bootstrap* paramétrico (Goldman 1993), não-paramétrico (Hasegawa et al. 1988) ou RELL (Kishino et al. 1990) para comparar a adequação de filogenias discrepantes. Isso fragiliza sua praticidade para enormes conjuntos de dados, cada vez mais empregados na era da Filogenômica. O ED se diferencia por obter a distribuição de Δln*L*s mais diretamente, via aproximação normal, mas não é o primeiro teste a lançar mão desse recurso. Hoje, a maior parte dos programas que implementam o KH utilizam métodos de reamostragem, mas em seu *paper* de introdução (Kishino & Hasegawa 1989) e sua primeira implementação em *software* (PHYLIP), esse teste não só pressupunha a normalidade assintótica de Δln*L*s entre modelos não-hierárquicos, como derivava a variância de sua distribuição a partir da variância dos Δln*L*s por sítio, calculados com o alinhamento original.

Contudo, em seus estudos das distribuições assintóticas de Δln*L*, Vuong (1989) demonstra que sua normalidade entre modelos não-hierárquicos se mantém apenas na condição de "igual proximidade" - medida pela divergência de Kullback-Leibler - entre eles e o modelo verdadeiro; perdendo simetria ao pender para valores mais positivos ou negativos, caso um dos modelos esteja mais próximo que o outro. O fato de que, com dados empíricos, não é possível provar a equidistância entre duas filogenias sub-ótimas e a geradora (desconhecida) dos dados faz com que o pressuposto de normalidade imponha, implicitamente, o pressuposto de equidistância ao KH. Consequentemente, na eventual violação dessa condição para as árvores testadas, não há qualquer evidência de que a variância derivada dos Δln*L*s por sítio entre elas equivaleria à variância da distribuição (normal) de Δln*L*s em um cenário no qual a equidistância se sustentasse e a esperança de Δln*L* fosse igual a 0 (hipótese nula do KH). A proposta fundamental do ED é nada mais do que corrigir esse problema, aproximando a variância de Δln*L* na condição de equidistância por uma via independente das verossimilhanças das árvores testadas.

Os resultados obtidos aqui oferecem um prognóstico promissor para essa alternativa. Lançando mão de simulações de seqüências, rearranjos topológicos e uma medida de distância topológica sensível a comprimentos de ramos, pude verificar a existência de

variáveis que influenciam o desvio padrão da distribuição de forma previsível, quando na condição de equidistância. Isso indica que, apesar da complexidade dos modelos probabilísticos empregados em filogenética, ao menos em condições especificas há meios para aproximar a forma da distribuição de Δln*L*s em menos tempo e com menos recursos computacionais. No entanto, antes que o modelo regredido possa ser validado para aplicações mais gerais, em estudos com outros grupos de táxons, é necessário discutir a relevância de cada variável já incluída no modelo e considerar possíveis aprimoramentos.

5.1.1 – Pelo número de sítios

Dado que cada sítio de uma matriz de dados é interpretado, normalmente, como uma observação indepentente (Felsenstein, 1981), a relação negativa entre o número de sítios e o desvio padrão de Δln*L* (**Figura 22b**) segue como consequência desse *status*. O mesmo conceito apresentado na introdução se aplica aqui: quanto maior é o número de observações na amostra, menor é o erro randômico (ou amostral) associado à estimação de cada parâmetro no modelo probabilístico. Da mesma forma, quanto mais extenso é o alinhamento, menor tende a ser o erro na estimação das taxas de substituição relativas, da frequência de equilíbrio para cada base, dos comprimentos de ramos nas filogenias e demais.

A razão pela qual esse fenômeno parece refletir diretamente na redução da variância da distribuição de Δln*L* pode estar associada a uma menor variação das verossimilhanças calculadas para as diferentes árvores, que, por sua vez, resultaria da maior precisão na estimação de seus parâmetros. Assim, valores de verossimilhança que variam menos erraticamente tenderiam a apresentar diferenças menores entre si, mesmo ao pontuar árvores com resoluções discrepantes.

É esperado que, à medida que o número de sítios tenda a infinito, o erro amostral tenda a 0 (Yang, 2006:186). Contudo, o fato da relação observada ter sido não-linear reitera que, ao menos em filogenética, a precisão das estimações aumenta mais rapidamente que o número de observações; ou seja, as estimativas dadas múltiplas amostras de tamanho crescente convergiriam para o mesmo valor bem antes de ser necessário um número infinito de sítios. Com mais tempo ou mais recursos computacionais, seria desejável reproduzir o experimento realizado aqui com tamanhos amostrais que fossem além de 2000 sítios, até atingir-se tal ponto de convergência. Isso validaria a precisão do coeficiente regredido para qualquer tamanho de alinhamento ao qual o modelo eventualmente fosse aplicado.

Todavia, voltando a outro conceito apresentado na introdução, ao utilizar dados moleculares empíricos, especialmente em Filogenômica, o aumento do tamanho da amostra

não necessariamente resulta na redução do erro total de estimação, já que a heterogeneidade de sinal entre genes pode violar o modelo - de substituição e/ou a própria filogenia - de múltiplas formas. Considerado isto e que, no experimento realizado aqui, cada alinhamento foi simulado a partir de uma única árvore, um teste topológico com o modelo regredido seria mais seguramente aplicado em escala gene-específica ou no máximo a grupos de genes que indiquem compartilhar a mesma filogenia.

5.1.2 – Pela proporção de *gaps*

Cada caractere correspondente a um *gap* (-) em um dado sítio do alinhamento é interpretado, pela maioria dos programas de inferência filogenética (RAxML, PhyML, IQ-tree, Mr. Bayes, PAML, BEAST, etc.) não como resultado da inserção ou deleção de uma base, mas como um caractere indeterminado. Em outras palavras, é compreendido como um pedaço de informação desconhecida, podendo corresponder a um A, um C, um T ou um G. Assim, a relação positiva entre a proporção de *gaps* no alinhamento e o desvio padrão de ΔlnL (**Figura 22c**) parece não ser nada mais que uma consequência numérica disso.

Mais detalhadamente, na presença de qualquer ambiguidade em uma das bases nas duas extremidades de um ramo (seja em um nó ancestral ou em um terminal da árvore – ver **Figura 6**), a probabilidade desse ramo será a soma das probabilidades que teria dada cada resolução possível para a base ambígua (A, C, T ou G) (Felsenstein, 1981). Por conseguinte, qualquer ramo terminal que leve a um *gap* terá caracteres indeterminados em ambas extremidades (no terminal e no ancestral) e probabilidade igual a 1. Como a probabilidade da árvore dado um sítio (ou verossimilhança por sítio) é o produto das probabilidades de todos os seus ramos e sua verossimilhança total, por vez, é o produto das verossimilhanças por sítio; quanto maior for a proporção de *gaps* no alinhamento, maior será o número de ramos com probabilidade 1, maiores serão as verossimilhanças por sítio e também a verossimilhança total.

Em um caso extremo, em que todos os caracteres fossem indeterminados, tal alinhamento não conteria qualquer informação filogenética e pontuaria verossimilhança máxima para qualquer árvore ($L = 1$ ou $\ln L = 0$). Em outros cenários mais realistas, à médida que a proporção de *gaps* no alinhamento aumenta, converge-se para a mesma situação; árvores discrepantes pontuando verossimilhanças distintas, mas que tendem a diferenças cada vez menores entre si, resultando, enfim, no efeito observado de redução do desvio padrão de $\Delta \ln L$.

Outros pesquisadores já demostraram, por outro lado, que caracteres ambíguos podem atuar reduzindo a acurácia das inferências filogenéticas (Wiens, 1998) e até gerar sinal não-filogenético, aumentando a verossimilhança de árvores com partições específicas, no eventual insucesso em contemplar a variação de taxas entre sítios (Lemmon et al. 2009). Apesar desses fenômenos potencialmente afetarem a variação de Δln*L* positivamente, o efeito observado aqui, com modelos de substituição corretamente especificados, foi o oposto.

5.1.3 – Pelo número de ramos

Um poderia concluir que, como mais sítios aumentam a precisão das estimativas, reduzindo a variação de Δln*L*, mais sequências no alinhamento apresentariam o mesmo efeito, tornando contraintuitiva a relação positiva observada entre essa variável e o desvio padrão de ΔlnL (**Figura 22d**). Na verdade, ao menos com árvores totalmente dicotômicas, mais sequências implicam maior número de ramos e mais parâmetros (comprimentos dos ramos) a serem estimados; ou seja, mais erro amostral (Yang 2006:37). O acúmulo de erro associado à estimação de cada parâmetro nas árvores com mais ramos pode ter causado uma maior variação nas verossimilhanças e, por consequência, o relativo aumento na variação de Δln*L*.

Entretanto, dentre as variáveis que apresentaram alguma relação com o desvio padrão de Δln*L*, o número de ramos foi a que ofereceu maior dificuldade de linearização. Mesmo após os esforços para garantir os pressupostos da regressão múltipla via transformações das variáveis, nota-se que a variação dos desvios padrões de Δln*L* pela quantidade de ramos na árvore foi a menos linear (**Figura 23d**), apesar de suficientemente monotônica. A possível explicação para isso é que exista outro fator associado à estrutura da árvore que influencie o erro na estimação dos comprimentos de ramo de forma diferenciada em cada filogenia.

A variável negligenciada poderia ter relação com a própria topologia ou com os tamanhos relativos dos ramos na árvore, podendo estar associada a dois dos fenômenos descritos por Schwartz & Muller (2010): (1) a precisão das estimativas de comprimento de ramo via ML é baixa quando o comprimento verdadeiro é muito curto e torna a cair sucessivamente para comprimentos verdadeiros superiores a 0,6 s/s e (2) quanto maior a profundidade do ramo na árvore (número de ramos entre ele e o terminal mais próximo), ainda mais imprecisas são as estimativas de seu tamanho. É imprescindível, para um futuro aprimoramento do modelo, que ainda seja avaliada a possível influência desses fatores sobre a distribuição de Δln*L*.

5.1.4 – Pelas distâncias topológicas

Quanto maior o erro associado à estimação dos comprimentos de ramo em árvores discrepantes, também espera-se que maior seja a distância topológica entre elas, desde que a medida leve em conta as diferenças entre seus comprimentos de ramos. Mesmo assim, foi surpreendente a relação positiva encontrada entre a distância topológica entre árvores igualmente distantes da verdadeira e a o desvio padrão de $\Delta\ln L$ (**Figura 22a**). Esperaria-se, naturalmente, ver uma relação positiva ao comparar, por exemplo, a árvore verdadeira a qualquer outra; conforme a segunda se afasta dos parâmetros verdadeiros no espaço paramétrico multidimencional, aumenta a distância topológica em relação à verdadeira, tendendo a obter diferenças de verossimilhança progressivamente maiores, aumentando a assimetria e o desvio da distribuição de $\Delta\ln L$ (ver Vuong (1989) para a descrição desse efeito em modelos lineares).

No entanto, é preciso levar em conta a variação irregular da verossimilhança ao longo do espaço topológico; a mesma que leva à formação de múltiplos picos de verossimilhança menores que a máxima verossimilhança global (ótimos locais) (Chor et al. 2000). Tais irregularidades poderiam levar ao aumento da variação de $\Delta\ln L$ entre topologias progressivamente distantes, mesmo que equidistantes da verdadeira. O fato de que o aumento da distância entre elas resulta, eventualmente, no aumento da distância em relação à verdadeira pode contribuir também; se de alguma forma o espaço topológico tornar-se mais irregular em regiões mais afastadas da verdadeira. Esta última possibilidade também explicaria o aumento da dispersão de $\Delta\ln L$ com o distanciamento das árvores comparadas em relação à verdadeira (**Figura 19f**). Se este for o caso, o efeito visto da distância entre alternativas (**Figura 19e**) seria mera consequência do aumento de sua distância para a verdadeira, portanto a última deveria ser considerada diretamente.

Com dados empíricos, não seria possível acessar a distância entre as árvores comparadas e a verdadeira (desconhecida). Entretanto, uma abordagem alternativa poderia nos permitir aproximá-la. Por exemplo, um teste indireto da condição de equisitância poderia determinar se a assimetria da distribuição amostral de $\Delta\ln L$ é significativamente diferente de 0. Então, se não rejeitasse a hipótese de simetria – e, indiretamente, de equidistância - um modelo de regressão similar ao que foi gerado aqui, porém incluindo a distância entre as árvores comparadas e a verdadeira, poderia aproximar esta variável a partir do desvio padrão da distribuição (e das demais variáveis já incluídas). A óbvia consequência de tal método, contudo, seria o sacrifício da economia computacional que se atinge ao dispensar

procedimentos de reamostragem e sua nulidade nos casos em que a distribuição amostral fosse assimétrica.

Enquanto forçados a negligenciar a distância topológica até a árvore verdadeira, seria interessante compor cenários (via rearranjos topológicos e simulações) que possibilitassem dissociar o efeito das duas distância sobre a distribuição de $\Delta\ln L$ e analisar até que nível tal negligência penaliza a precisão do modelo de regressão. Além disso, seria desejável verificar como regiões, no espaço topológico, que apresentam árvores de verossimilhança idêntica (terraços) podem afetar a distribuição. Próximo a esses terraços, espera-se que ocorra uma redução abrupta do desvio padrão de $\Delta\ln L$, independente de quaisquer distâncias topológicas. Por outro lado, as condições necessárias para a formação de terraços só foram vistas, até hoje, em análises multi-*locus* particionadas (Sanderson et al. 2011), podendo não vir a ser uma real preocupação em análises *locus*-específicas.

5.1.5 – Outras variáveis

Com os resultados obtidos pude caracterizar, na condição de equidistância, a influência de múltiplas variáveis sobre o desvio padrão de $\Delta\ln L$, com alto coeficiente de determinação. No entanto, por motivos de viabilidade computacional, todas as simulações foram limitadas a árvores binárias e a um único modelo de substituição (GTR com taxas heterogêneas entre sítios). Apesar dessa combinação ser uma das mais empregados para inferência filogenética, raramente atingindo níveis superiores de complexidade em análises de árvores de genes por ML, é preciso atentar à possível influência da parametrização do modelo sobre a precisão das estimativas.

Como discutido anteriormente, espera-se que o número de parâmetros no modelo evolutivo afete a quantidade de erro amostral gerado, aumentando ou reduzindo a variação de $\Delta\ln L$ (Yang 2006:37). Ainda, dependendo de qual parâmetro livre é reduzido ou adicionado, é possível que o desvio padrão da distribuição seja afetado em diferentes intensidades. Isso torna necessária a reprodução do experimento com casos específicos do GTR (HKY, TN93, F81 , K2P, JC, etc) e, assim que houver *software* capaz de implementar modelos mais complexos em inferências por ML, simulações que comportem diferentes taxas (Kolaczkowski & Thornton, 2008), matrizes de substituição (Blanquart & Lartillot, 2008) e vetores de composição (Foster, 2004) ao longo da árvore.

Finalmente, será necessário ainda delimitar o poder de rejeição do teste de equidistância. Lançando mão de mais simulações, será possível verificar sua proporção de rejeição da hipótese nula entre árvores de distâncias progressivamente diferentes em relação à

verdadeira e, assim, determinar a partir de que nível de violação da equidistância o teste é capaz de rejeitá-la com 95% de probabilidade. Um modelo de regressão de propriedades mais desejáveis deve equilibrar generalidade, sendo aplicável a qualquer alinhamento de sequências de um mesmo gene, e capacidade de controle de suas taxas de erro (tipo I e II) com diferença mínima de distâncias entre as árvores comparadas e a árvore geradora.

5.2 – Estimativas pontuais para a raiz de Placentalia

Não surpreendentemente, as três hipóteses mais aceitas na bibliografia para a posição da raiz de Placentalia lideraram conjuntamente em frequência entre as árvores de gene inferidas. Esse resultado reitera a dificuldade em solucionar o icônico dilema. Sendo cada vez menos incontroverso que um rápido processo de surgimento das superordens, dirigido por sucessivos eventos de vicariância na fragmentação da Pangea e Gondwana, tenha causado amostragem incompleta de linhagens gênicas entre placentários (Nishihara et al. 2009; Hallström & Janke 2010), alguns chegam a aventar a impossibilidade de uma solução dicotômica para essa parte da história do grupo (Hallström & Janke 2010). Em contrapartida, outros apontam que problemas de natureza metodológica, a exemplo de viéses de composição nos dados (Romiguier et al. 2013) e limitações nos modelos de substituição tradicionais (Morgan et al. 2013), podem gerar resoluções artefatuais, impedindo o consenso entre estudos diversos.

5.2.1 – Conteúdo GC, tamanho dos *loci* e acúmulo de substituições

Como visto anteriormente, Romiguier e colaboradores (2013) advogam pelo uso de *loci* ricos em bases fracas (A e T) como prevenção de sinal não-filogenético. O argumento se sustenta em evidências de que, após eventos de recombinação, a maquinaria de correção genômica tende a corrigir *mismatches* com G e C mais frequentemente que com A ou T, tornando o conteúdo GC um indicador de *hot-spots* de recombinação em determinadas regiões genômicas (Lartillot, 2013). Baseados nisso, os autores opinam que *loci* ricos em GC seriam mais homoplásticos, devido ao dinamismo desses *hot-spots.* Relatam, também, que, em suas análises, houve maior tendência desses *loci* a recuperar Atlantogenata e Boreoeutheria como resolução para a raiz de Placentalia (Romiguier et al. 2013), o que faria dessa resolução um artefato. Entretanto, os resultados obtidos aqui falharam em demonstrar correlação entre a proporção GC e qualquer uma das três principais resoluções (**Figura 25**), levantando a possibilidade da tendência relatada pelos autores ser, a própria, um artefato.

Mesmo na possibilidade da relação vista por Romiguier e coautores (2013) ter sido efeito indireto de tamanho diferenciado entre *loci;* no conjunto de dados analisado aqui também não foi vista associação entre os maiores genes e qualquer uma das principais resoluções obtidas para a raiz (**Figura 26**). O único resultado reproduzido com sucesso foram as distâncias relativamente superiores entre o nó MRCA de Placentalia e os nós de Afrotheria e Xenarthra nas árvores que recuperam Atlantogenatha (**Figura 27**). Apesar de isso sugerir um aumento da probabilidade de atração de ramos longos, é preciso atentar ao fato de que um maior acúmulo de substituições não necessariamente resulta em saturação; para que a condição se afirme, é necessário não só que ocorra um grande número de substituições em ambas as linhagens, mas que o modelo evolutivo falhe em estimar o total dessas substituições, consequentemente tomando homoplasias por homologias.

Ademais, Romiguier e colaboradores (2013) reiteram a hipótese de atração de ramos longos ao relatarem maior "incongruência topológica" nas árvores recuperadas de *loci* ricos em GC. No entanto, a medida de congruência que utilizam é o nível de concordância entre as relações estimadas por cada grupo de genes (dos mais ricos em AT até os mais ricos em GC) e as de uma árvore referência, baseada nas inferências de dois trabalhos específicos (Springer et al. 2004; Meredith et al. 2011) e reduzida a politomias em seus pontos alegadamente mais controversos. Como outros pesquisadores, acredito que esse critério penaliza a medida de incongruência por depender das relações consideradas incontroversas pelos autores. Medidas mais apropriadas, como a distância topológica média entre árvores estimadas, também não revelam relação entre o conteúdo GC e o consenso topológico na árvore de Mammalia (Filipe Romero, não publicado).

Em suma, por si só, o acúmulo desproporcional de substituições entre o MRCA de Placentalia e Afrotheria/Xenathra não basta como evidência de atração de ramos longos; sendo necessário, adicionalmente, determinar se a inferência de Atlantogenata resulta de informação filogenética ou não-filogenética. Morgan e colaboradores (2013) advogaram pelo uso de modelos mais complexos para separar "o joio do trigo" em Placentalia.

5.2.2 – Parametrização das inferências

É cada vez mais reconhecido o potencial dos chamados modelos 'sítio-heterogêneos' em evitar artefatos de atração de ramos longos (Lartillot et al. 2007; Foster et al. 2009). Acomodando diferentes frequências de equilíbrio ao longo do alinhamento, modelos sítio-heterogêneos, como o CAT para substituições de aminoácidos, evitam a subestimação de eventos de substituição em sítios de evolução diferenciada, principalmente quando há viés

composicional associado (Lartillot et al. 2007). Adicionalmente, modelos 'tempo-heterogêneos', sensíveis a heterotaquia (ver Lopez et al. 2002), podem ainda contemplar variações nas taxas de substituição entre linhagens (ramos) na árvore (Kolaczkowski & Thornton 2008; Blanquart & Lartillot 2008). Por outro lado, o emprego desses modelos exige que a maior parametrização seja compensada por um grande número de sítios no alinhamento (Quang et al. 2008), para garantir a precisão das estimativas apesar de sua tendência a maior acúmulo de erro amostral.

As análises concatenadas realizadas aqui permitiram avaliar a influência dessa parametrização progressiva sobre a resolução da raiz de Placentalia. Apesar do particionamento das análises não caracterizar propriamente a aplicação de um modelo sítio-heterogêneo, de maneira similar a um, permite razoável acomodação da heterogeneidade no processo evolutivo entre as partições definidas; sendo também computacionalmente mais viável (Lanfear et al. 2014). Adicionalmente, a estimação dos tamanhos de ramos independentemente para cada partição é análoga à soma das verossimilhanças para múltiplos *sets* de comprimentos de ramos, realizada nos modelos tempo-heterogêneos, sensíveis a heterotaquia. Contudo, nem as análises particionadas com ramos proporcionais entre partições, nem as particionadas com ramos independentes recuperaram resolução diferente das análises não-particionadas, que permitiram variação apenas das taxas de substituição entre sítios (via distribuição gama).

A ausência de efeito da maior parametrização sobre a topologia recuperada, especialmente por deixar intacto o monofiletismo de Boreoeutheria e Atlantogenata, sustenta que o agrupamento de Xenarthra e Afrotheria pode resultar de sinal filogenético legitimo. Esses resultados corroboram, ainda, com o que se obtém ao aplicar modelos mais sofisticados de 'sítio-heterogeneidade' e 'tempo-heterogeneidade' a conjuntos de dados para Placentalia: Morgan e colaboradores (2013) também recuperaram a raiz entre Atlantogenata e Boreoeutheria com alto suporte ao aplicar um modelo CAT-GTR, além de modelos mistos com duas matrizes de substituição GTR e 4 vetores de composição. Ademais, ao reanalisar os conjuntos de dados de trabalhos anteriores (Hallström & Janke 2010; O'Leary et al. 2013; Romiguier et al. 2013) com modelos CAT-GTR, Tarver e coautores (2016) reverteram os resultados diversos desses trabalhos em favor da mesma hipótese; obtiveram baixo suporte de *bootstrap* para Atlantogenata apenas pelos dados de Romiguier e colaboradores (2013), mas recuperando, ainda assim, seu monofiletismo em detrimento de Afrotheria externo.

5.3 – Testes topológicos e a significância das estimativas

Apesar da raiz entre Boreoeutheria e Atlantogenata ter sido a resolução menos frequente - das três principais - entre as árvores de genes inferidas, aquelas que reproduziram essa posição apresentaram o menor nível de rejeição pelos testes topológicos aplicados. Não obstante, todas as três resoluções mais frequentes tiveram proporções de rejeição insignificantes (menos de 5% dos *loci*) quando confrontadas entre si (**Tabelas 1 e 2**); chegando a proporções pouco maiores apenas pelo AU, mas não passando de 6,2%; o que aponta uma equivalência de suporte estatístico para essas três hipóteses em praticamente todos os genes analisados.

Contudo, quando se compara cada uma das três às demais resoluções possíveis para as relações entre superordens, as proporções de rejeição da hipótese nula, em todos os restes, sobem vertiginosamente em detrimento das últimas; mas significativamente menos quando a alternativa contém Atlantogenata, mesmo que não preserve Boreoeutheria como grupo monofilético. Dado que essa diferença é ainda mais acentuada nos resultados do ED, seria interessante, ainda, verificar se o resultado se mantém após os devidos aprimoramentos do modelo de regressão. Mesmo assim, tendo sido concomitante entre todos os testes topológicos, acredito que esse resultado aponte a presença de um sinal críptico para o monofiletismo de Atlantogenta em alguns genes que não recuperam o grupo em suas árvores ML, provavelmente por ter sido suplantado pelos sítios de sinal mais ambíguo nesses *loci*.

Em síntese, as análises concatenadas e os testes topológicos gene-específicos corroboraram para trazer à tona suporte estatístico para Atlantogenata, antes velado pelas estimativas das árvores de genes. Essas conclusões não dispensam a necessidade de se levar em conta a heterogeneidade de sinal entre genes, resultante de recombinação ou ILS; pelo contrário, destacam a importância da conjunção entre análises concatenadas adequadamente modeladas, testes topológicos para determinar o suporte estatístico de cada árvore de gene estimada e, só então, análises de coalescência multiespécies com aqueles genes que apresentarem sinal menos ambíguo – com os quais for possível rejeitar as hipóteses alternativas em questão.

Em futuros tratamentos do problema da raiz de Placentalia, é de suma importância, também, que se verifique o efeito do aumento da amostragem taxonômica para Xenarthra e Afrotheria sobre as resoluções; mesmo que a custo de uma redução na amostragem de *loci*. Apesar da maior disponibilidade de sequências atualmente, um problema comum de muitos trabalhos que tratam o problema da raiz de Placentalia - este incluso - está na desproporção da amostragem de táxons entre as superordens (McCormack et al. 2012; Romiguier et al. 2013;

Morgan et al. 2013; Tarver et al. 2016). Árvores mais densas podem tender a resoluções mais heterogêneas por efeito de amostragem incompleta de linhagens (Rosenberg 2002); por outro lado, em conjunto com as demais estratégias discutidas, uma melhor representação das superordens que compõem Atlantogenata pode quebrar os ramos longos de suas linhagens, ajudando a definir se o agrupamento de fato não é artefato de saturação.

6 – Conclusões

Na condição de equidistância entre as árvores comparadas e a verdadeira e na correta especificação do modelo de substituição, o desvio padrão da distribuição amostral de $\Delta \ln L$ pode ser aproximado sem qualquer informação sobre a verossimilhança das árvores comparadas. Isso se deve, principalmente, à contribuição das distâncias topológicas entre elas para a variação da estatística $\Delta \ln L$, como também, em menor escala, do número de ramos nas árvores, do número de sítios no alinhamento e da proporção de *missing data*. A aplicação de um modelo de regressão dessas variáveis para testar a hipótese de equidistância com dados empíricos corroborou essa previsibilidade ao mostrar performance, em geral, similar ao de outros testes topológicos tradicionais. O novo teste possibilitou, também, expressiva vantagem computacional pela aproximação normal, prenunciando grande utilidade para análises de viés topológico em conjuntos de dados genômicos.

Entretanto, múltiplos aprimoramentos do modelo de regressão ainda são necessários para garantir sua generalidade e aplicabilidade a outros conjuntos de dados; bem como uma avaliação de suas taxas de erro do tipo I e tipo II, para verificar se mantém acurácia diante dessa sofisticação. É preciso incluir, por exemplo, o possível efeito da quantidade e do tipo de cada parâmetro livre no modelo de substituição, já que aqui somente o modelo GTR foi empregado. Ainda mais importante para a viabilidade do modelo de regressão, seria determinar até que ponto a variação na distância entre as árvores comparadas pode aproximar a variação na distância em relação à verdadeira; que também tem efeito sobre a distribuição de $\Delta \ln L$, mas não pode ser acessada.

O estudo de caso com a raiz de Placentalia também revelou, através dos testes topológicos, a existência de sinal críptico para Atlantogenata entre os genes analisados; isto é, suporte estatístico insuficiente para a rejeição de árvores com esse grupo, por genes que não o recuperaram em suas estimativas pontuais de ML. O fato de todos os testes topológicos terem reproduzido esse fenômeno, apesar de Atlantogenata + Boreoeutheria ter sido a resolução menos frequente entre as árvores de gene estimadas - quando comparada a Afrotheria ou Xenarthra externos - reitera a utilidade dos testes de topologia ao se empregar métodos que tratam essas árvores como observações. Nesses casos, é recomendada a utilização dos testes como critério de seleção dos genes com sinal mais decisivo para cada árvore estimada. Futuramente, uma versão sofisticada do modelo empregado aqui pode viabilizar esse procedimento.

Referências

Aken, B. L., Ayling, S., Barrell, D., Clarke, L., Curwen, V., Fairley, S., … Searle, S. M. J. (2016). The Ensembl Gene Annotation System. Database : The Journal of Biological Databases and Curation, 2016, baw093. https://doi.org/10.1093/database/baw093

Anisimova, M., & Gascuel, O. (2006). Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. Syst Biol, 55(4), 539–552. https://doi.org/T808388N86673K61

Bergsten, J. (2005). A review of long-branch attraction, 21, 163–193. https://doi.org/10.1111/j.1096-0031.2005.00059.x

Blanquart, S., & Lartillot, N. (2008). A site- and time-heterogeneous model of amino acid replacement. Molecular Biology and Evolution, 25(5), 842–858. https://doi.org/10.1093/molbev/msn018

Burnham, K. P., & Anderson, D. R. (2004). Multimodel Inference. Sociological Methods & Research, 33(2), 261–304. https://doi.org/10.1177/0049124104268644

Chor, B., Hendy, M. D., Holland, B. R., & Penny, D. (2000). Multiple maxima of likelihood in phylogenetic trees: an analytic approach. Molecular Biology and Evolution, 17(10), 1529–41.

Cox, S. R. (1961). Tests of separate families of hypotheses. Proc. 4th Berkeley Symp. Math. Stat. Prob. 1:105–23

Degnan, J. H., & Rosenberg, N. A. (2009). Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends in Ecology and Evolution, 24(6), 332–340. https://doi.org/10.1016/j.tree.2009.01.009

Dobzhansky, T. (1941). Genetics and the origin of species. Comlumbia University Press.

Douady, C. J., Chatelier, P. I., Madsen, O., De Jong, W. W., Catzeflis, F., Springer, M. S., &

Stanhope, M. J. (2002). Molecular phylogenetic evidence confirming the Eulipotyphla concept and in support of hedgehogs as the sister group to shrews. Molecular Phylogenetics and Evolution, 25(1), 200–209. https://doi.org/10.1016/S1055-7903(02)00232-4

Douzery, E. J. P., Scornavacca, C., Romiguier, J., Belkhir, K., Galtier, N., Delsuc, F., & Ranwez, V. (2014). OrthoMaM v8: A database of orthologous exons and coding sequences for comparative genomics in mammals. Molecular Biology and Evolution, 31(7), 1923–1928. https://doi.org/10.1093/molbev/msu132

Edwards, A. W. F., & Cavalli-Sforza, L. L. (1964). Reconstruction of evolutionary trees. Phenetic and Phylogenetic Classification, 6(6), 67–76.

Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. The Annals of Statistics, 7(1), 1–26. https://doi.org/10.1214/aos/1176344552

Efron, B., Halloran, E., & Holmes, S. (1996). Bootstrap confidence levels for phylogenetic trees. TL - 93. Proceedings of the National Academy of Sciences of the United States of America, 93, 7085–70904. https://doi.org/10.1073/pnas.93.23.13429

Felsenstein, J & Kishino, H. (1993). Is there something wrong with the bootstrap on phylogenies? A reply to Hillis and Bull. Systematic Biology, 42(2), 193–200.

Felsenstein, J. (1978). Cases in which parsimony or compatibility methods will be positively misleading. Systematic Biology, 27(4), 401–410.

Felsenstein, J. (1985). Confidence limits on phylogenies: an approach using the bootstrap. Evolution, 39(4), 783–791. https://doi.org/10.2307/2408678

Felsenstein, J. (1988). Phylogenies from molecular sequences: inference and reliability. Annual Review of Genetics, 22, 521–565.

Felsenstein, J. (1989). PHYLIP - Phylogeny Inference Package (Version 3.2). Cladistics, 5, 164-166.

Fienberg, S. E. (2006). When did Bayesian inference become “Bayesian"? Bayesian Analysis, (1), 1–41. https://doi.org/http://dx.doi.org/10.1214/06-BA101

Fisher, R. A. (1918). The Correlation between relatives on the supposition of Mendelian Inheritance. Transactions of the Royal Society of Edinburgh, 52:399-433.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. Philosophical Transactions of the Royal Society of London, 222, 309–368.

Fisher, R. A. (1930). The Genetical Theory of Natural Selection. Oxford Univerity Press. https://doi.org/10.1038/158453a0

Fitch, W. & Margoliash, E. (1967). Construction of Phylogenetic Trees. Science, 155(3760), 279–284.

Fitch, W. M. (1970). Distinguishing Homologous from Analogous Proteins. Systematic Biology, 19(2), 99–113. https://doi.org/10.2307/2412448

Fletcher, W., & Yang, Z. (2009). INDELible: A flexible simulator of biological sequence evolution. Molecular Biology and Evolution, 26(8), 1879–1888. https://doi.org/10.1093/molbev/msp098

Foley N. M., Springer M. S., Teeling EC (2016) Data from: Mammal madness: is the mammal tree of life not yet resolved? Dryad Digital Repository. http://dx.doi.org/10.5061/dryad.7309c

Foley, N. M., Springer, M. S., & Teeling, E. C. (2016). Mammal madness: is the mammal tree of life not yet resolved? Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 371(1699), 3667–3679. https://doi.org/10.1098/rstb.2015.0140

Foster, P. (2004). Modeling Compositional Heterogeneity. Systematic Biology, 53(3), 485–495. https://doi.org/10.1080/10635150490445779

Foster, P. G., & Hickey, D. a. (1999). Compositional bias may affect both DNA-based and protein-based phylogenetic reconstructions. Journal of Molecular Evolution, 48(3), 284–290. https://doi.org/10.1007/PL00006471

Gatesy, J., Milinkovitch, M., Waddell, V., & Stanhope, M. (1999). Stability of Cladistic Relationships between Cetacea and Higher-Level Artiodactyl Taxa. Syst. Biol, 48(1), 6–20. https://doi.org/10.1080/106351599260409

Ghasemi, A., & Zahediasl, S. (2012). Normality tests for statistical analysis: A guide for non-statisticians. International Journal of Endocrinology and Metabolism, 10(2), 486–489. https://doi.org/10.5812/ijem.3505

Goldman, N. (1993). Statistical Tests of Models of DNA Substitution. Journal of Molecular Evolution, 36, 182–198.

Goldman, N., Anderson, J. P., & Rodrigo, a G. (2000). Likelihood-based tests of topologies in phylogenetics. Systematic Biology, 49(4), 652–670. https://doi.org/10.1080/106351500750049752

Goodman M., Czelusniak J., Moore G. W., Romero-Herrera A.E., & Matsuda G. (1979). Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed from globin sequences. Systematic Zoology, 28, 132–163.

Graunt, J. (2013). 2. Natural and Political Observations Mentioned in a Fallowing Index, and Made Upon the Bills of Mortality, 1964(1662). https://doi.org/10.1007/978-3-642-35858-6

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). Multivariate Data Analysis. Vectors. https://doi.org/10.1016/j.ijpharm.2011.02.019

Hallström, B. M., & Janke, A. (2010). Mammalian evolution may not be strictly bifurcating. Molecular Biology and Evolution, 27(12), 2804–2816. https://doi.org/10.1093/molbev/msq166

Hillis, D. M., Mable, B. K., & Moritz, C. M. (1996). Applications of molecular systematics: the state of the field and a look to the future. 515-543 em Molecular systematics (D. M.Hillis, C. Moritz, and B. K. Mable, eds.). Sinauer, Sunderland, Massachusetts.

Hillis, D.M. & Bull, J. J. (1993). An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. Systematic Biology, 42(2), 182–192.

Huelsenbeck, J. P. (1995). The robustness of two phylogenetic methods: four-taxon simulations reveal a slight superiority of maximum likelihood over neighbor joining. Molecular Biology and Evolution, 12(5), 843–849.

Huensenback, J. P., & Crandall, K. A. (1997). Phylogeny estimation and hypothesis testing using Maximum Likelihood. Annual Review of Ecology, 28(437–466).

Hurvich, C. M., & Tsai, C.-L. (1993). a Corrected Akaike Information Criterion for Vector Autoregressive Model Selection. Journal of Time Series Analysis, 14(3), 271–279. https://doi.org/10.1111/j.1467-9892.1993.tb00144.x

Jeffroy, O., Brinkmann, H., Delsuc, F., & Philippe, H. (2006). Phylogenomics: the beginning of incongruence? Trends in Genetics, 22(4), 225–231. https://doi.org/10.1016/j.tig.2006.02.003

John, K. S. (2017). Review paper: The shape of phylogenetic treespace. Systematic Biology, 66(1), e83–e94. https://doi.org/10.1093/sysbio/syw025

Jukes, T. H., and C. R. Cantor. 1969. Evolution of protein molecules. 21–123 em Mammalian protein metabolism, volume III (H. N. Munro et al.). Academic Press, New York.

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Molecular Biology and Evolution, 30(4), 772–780. https://doi.org/10.1093/molbev/mst010

Kingman, J. F. C. (1982). The coalescent. Stochastic Processes and Their Applications, 13(3), 235–248. https://doi.org/10.1016/0304-4149(82)90011-4

Kishino, H., & Hasegawa, M. (1989). Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. Journal of Molecular Evolution, 29(2), 170–179. https://doi.org/10.1007/BF02100115

Kishino, H., Miyata, T., & Hasegawa, M. (1990). Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. Journal of Molecular Evolution, 31(2), 151–160. https://doi.org/10.1007/BF02109483

Kolaczkowski, B., & Thornton, J. W. (2008). A mixed branch length model of heterotachy improves phylogenetic accuracy. Molecular Biology and Evolution, 25(6), 1054–1066. https://doi.org/10.1093/molbev/msn042

Kuhner, M. K., & Felsenstein, J. (1994). A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Mol Biol Evol, 11(3), 459–468.

Kullback, S., Leibler, R. A. (1951). On Informatio and Sufficiency. The Annals of Mathematical Statistics, 22(1), 79–86.

Lanfear, R., Calcott, B., Kainer, D., Mayer, C., & Stamatakis, A. (2014). Selecting optimal partitioning schemes for phylogenomic datasets. BMC Evolutionary Biology, 14(1), 82. https://doi.org/10.1186/1471-2148-14-82

Lartillot, N., & Philippe, H. (2004). A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Molecular Biology and Evolution, 21(6), 1095–1109. https://doi.org/10.1093/molbev/msh112

Le Quesne, W. J. (1969). A method of selection of characters in numerical taxonomy. Systematic Zoology, 18(2), 201. https://doi.org/10.2307/2412604

O'Leary, M. a, Bloch, J. I., Flynn, J. J., Gaudin, T. J., Giallombardo, A., Giannini, N. P., … Cirranello, A. L. (2013). The placental mammal ancestor and the post-K-Pg radiation of placentals. Science (New York, N.Y.), 339(6120), 662–7. https://doi.org/10.1126/science.1229237

Leliaert, F., Verbruggen, H., Vanormelingen, P., Steen, F., López-Bautista, J. M., Zuccarello, G. C., & De Clerck, O. (2014). DNA-based species delimitation in algae. European Journal of Phycology, 49(2), 179–196. https://doi.org/10.1080/09670262.2014.904524

Lemmon, A. R., Brown, J. M., Stanger-Hall, K., & Lemmon, E. M. (2009). The effect of ambiguous data on phylogenetic estimates obtained by maximum likelihood and bayesian inference. Systematic Biology, 58(1), 130–145. https://doi.org/10.1093/sysbio/syp017

Li, W. H., Tanimura, M., & Sharp, P. M. (1987). An evaluation of the molecular clock hypothesis using mammalian DNA sequences. Journal of Molecular Evolution, 25(4), 330–342. https://doi.org/10.1007/BF02603118

Lopez, P., Casane, D., & Philippe, H. (2002). Heterotachy, an important process of protein evolution. Molecular Biology and Evolution, 19(September 2017), 1–7. https://doi.org/10.1093/oxfordjournals.molbev.a003973

Luo, Z. X., Cifelli, R. L., & Kielan-Jaworowska, Z. (2001). Dual origin of tribosphenic mammals. Nature, 409(6816), 53–57. https://doi.org/10.1038/35051023

Maddison, W. P. (1997). Gene trees in species trees. Systematic Biology, 46(3), 523–536. https://doi.org/10.1017/CBO9781107415324.004

Mayr, Ernst (1942). Systematics and the Origin of Species from the Viewpoint of a Zoologist. Harvard University Press, Cambridge, Massachusetts.

McCormack, J. E., Faircloth, B. C., Crawford, N. G., Gowaty, P. A., Brumfield, R. T., & Glenn, T. C. (2012). Ultraconserved elements are novel phylogenomic markers that resolve placental mammal phylogeny when combined with species-tree analysis. Genome Research, 22(4), 746–754. https://doi.org/10.1101/gr.125864.111

Meredith, R. W., Janečka, J. E., Gatesy, J., Ryder, O. a, Fisher, C. a, Teeling, E. C., … Murphy, W. J. (2011). Impacts of the Cretaceous Terrestrial Revolution and KPg extinction on mammal diversification. Science (New York, N.Y.), 334(6055), 521–4. https://doi.org/10.1126/science.1211028

Morgan, C. C., Foster, P. G., Webb, A. E., Pisani, D., & Mcinerney, J. O. (2013). Heterogeneous Models Place the Root of the Placental Mammal Phylogeny, 30(9), 2145–2156. https://doi.org/10.1093/molbev/mst117

Murphy, W. J., Eizirik, E., Johnson, W. E., Zhang, Y. P., Ryder, O. a, & O'Brien, S. J. (2001). Molecular phylogenetics and the origins of placental mammals. Nature, 409(6820), 614–618. https://doi.org/10.1038/35054550

Nakhleh, L. (2013). Computational approaches to species phylogeny inference and gene tree reconciliation. Trends in Ecology and Evolution, 28(12), 719–728. https://doi.org/10.1016/j.tree.2013.09.004

Neyman, J. Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. Philosophical Transactions of the Royal Society of London, 231, 289–337.

Nguyen, L., Schmidt, H. A., Haeseler, A. Von, & Minh, B. Q. (2014). IQ-TREE : A fast and effective stochastic algorithm for estimating maximum likelihood phylogenies, 1–19. https://doi.org/10.1093/molbev/msu300

Nichols, R. (2001). Gene trees and species trees are not the same. Trends in Ecology & Evolution, 16(7), 358–364. https://doi.org/10.1016/S0169-5347(01)02203-0

Nishihara, H., Maruyama, S., & Okada, N. (2009). Retroposon analysis and recent geological data suggest near-simultaneous divergence of the three superorders of mammals. Proceedings of the National Academy of Sciences of the United States of America, 106(13), 5235–5240. https://doi.org/10.1073/pnas.0809297106

Nishihara, H., Okada, N., & Hasegawa, M. (2007). Rooting the eutherian tree: the power and pitfalls of phylogenomics. Genome Biology, 8(9), R199. https://doi.org/10.1186/gb-2007-8-9-r199

Oliver, J. C. (2013). Microevolutionary processes generate phylogenomic discordance at ancient divergences. Evolution, 67(6), 1823–1830. https://doi.org/10.1111/evo.12047

Ota, R., Waddell, P. J., Hasegawa, M., Shimodaira, H., & Kishino, H. (2000). Appropriate likelihood ratio tests and marginal distributions for evolutionary tree models with constraints on parameters. Mol Biol Evol, 17(5), 798–803. https://doi.org/10.1093/oxfordjournals.molbev.a026358

Philippe, H., Brinkmann, H., Lavrov, D. V., Littlewood, D. T. J., Manuel, M., Wörheide, G., & Baurain, D. (2011). Resolving difficult phylogenetic questions: Why more sequences are not enough. PLoS Biology, 9(3). https://doi.org/10.1371/journal.pbio.1000602

Phillips, M. J., & Penny, D. (2003). The root of the mammalian tree inferred from whole mitochondrial genomes. Molecular Phylogenetics and Evolution, 28(2), 171–185. https://doi.org/10.1016/S1055-7903(03)00057-5

Phillips, M. J., McLenachan, P. a, Down, C., Gibb, G. C., & Penny, D. (2006). Combined mitochondrial and nuclear DNA sequences resolve the interrelations of the major Australasian marsupial radiations. Systematic Biology, 55(1), 122–137. https://doi.org/10.1080/10635150500481614

Posada, D. (2003). LETTERS jModelTest: Phylogenetic Model Averaging, 2001–2004. https://doi.org/10.1093/molbev/msn083

Quang, L. S., Gascuel, O., & Lartillot, N. (2008). Empirical profile mixture models for phylogenetic reconstruction. Bioinformatics, 24(20), 2317–2323. https://doi.org/10.1093/bioinformatics/btn445

Rannala, B., & Yang, Z. (2015). Efficient Bayesian species tree inference under the multi-species coalescent. Systematic Biology, 0(0), 1–21. https://doi.org/10.1093/sysbio/syw119

Ranwez, V., Harispe, S., Delsuc, F., & Douzery, E. J. P. (2011). MACSE: Multiple alignment of coding SEquences accounting for frameshifts and stop codons. PLoS ONE, 6(9). https://doi.org/10.1371/journal.pone.0022594

Robinson, D. . (1971). Comparison of labeled trees with valency three. Journal of Combinatorial Theory, Series B, 11(2), 105–119. https://doi.org/10.1016/0095-8956(71)90020-7

Romiguier, J., Ranwez, V., Delsuc, F., Galtier, N., & Douzery, E. J. P. (2013). Less is more in mammalian phylogenomics: AT-rich genes minimize tree conflicts and unravel the root of placental mammals. Molecular Biology and Evolution, 30(9), 2134–2144. https://doi.org/10.1093/molbev/mst116

Rosenberg, N. a. (2002). The probability of topological concordance of gene trees and species trees. Theoretical Population Biology, 61(2), 225–247. https://doi.org/10.1006/tpbi.2001.1568

Rothman, K. J., Greenland, S., & Associate, T. L. L. (2014). Modern Epidemiology, 3rd Edition. The Hastings Center Report, 44 Suppl 2, insidebackcover. https://doi.org/10.1002/hast.292

Sanderson, M. J., McMahon, M., & Steel, M. (2017). Terraces in Phylogenetic Tree Space. Science, 333(June 2011), 448–450. https://doi.org/10.1126/science.1206357

Schwartz, R. S., & Mueller, R. L. (2010). Branch length estimation and divergence dating: estimates of error in Bayesian and maximum likelihood frameworks. BMC Evol. Biol., 10, 1–21. https://doi.org/10.1186/1471-2148-10-5

Self, S., & Liang, K. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. J Am Stat Assoc, 82(398), 605–610. https://doi.org/10.1080/01621459.1987.10478472

Shimodaira, H. (2002). An Approximately Unbiased Test of Phylogenetic Tree Selection. Systematic Biology, 51(3), 492–508. https://doi.org/10.1080/10635150290069913

Shimodaira, H., & Hasegawa, M. (1999). Letter to the Editor: Multiple Comparisons of Log-Likelihoods with Applications to Phylogenetic Inference, Mol. Biol. Evol. 16:1114–1116.

Song, S., Liu, L., Edwards, S. V., & Wu, S. (2012). Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. Proceedings of the National Academy of Sciences, 109(37), 14942–14947. https://doi.org/10.1073/pnas.1211733109

Springer, M. S., Stanhope, M. J., Madsen, O., & De Jong, W. W. (2004). Molecules consolidate the placental mammal tree. Trends in Ecology and Evolution, 19(8), 430–438. https://doi.org/10.1016/j.tree.2004.05.006

Strimmer, K., & Rambaut, A. (2002). Inferring con é dence sets of possibly misspeci é ed gene trees, (September 2001). https://doi.org/10.1098/rspb.2001.1862

Swofford, D. L., Olsen G. J., Waddell P. J., D.M. Hillis (1996). Phylogenetic inference. 407–514 em Molecular systematics (D. M.Hillis, C. Moritz, and B. K. Mable, eds.). Sinauer, Sunderland, Massachusetts.

Tarver, J. E., Dos Reis, M., Mirarab, S., Moran, R. J., Parker, S., O'Reilly, J. E., … Pisani, D. (2016). The interrelationships of placental mammals and the limits of phylogenetic inference. Genome Biology and Evolution, 8(2), 330–344. https://doi.org/10.1093/gbe/evv261

Vallender, E. J. (2010). Bioinformatic approaches to identifying orthologs and assessing evolutionary relationships, 49(1), 50–55. https://doi.org/10.1016/j.ymeth.2009.05.010.

Van Dijk, E. L., Auger, H., Jaszczyszyn, Y., & Thermes, C. (2014). Ten years of next-generation sequencing technology. Trends in Genetics, 30(9). https://doi.org/10.1016/j.tig.2014.07.001

VanderPlas, J. (2014). Frequentism and Bayesianism: A Python-driven Primer. arXiv, astro-I(Scipy), 1–9.

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. Econometrica, 57(2), 307–333.

Waddell, P. J., & Steel, M. . (1997). General Time-Reversible Distances with Unequal Rates across Sites: Mixing Γ and Inverse Gaussian Distributions with Invariant Sites. Molecular Phylogenetics and Evolution, 8(3), 398–414. https://doi.org/10.1006/mpev.1997.0452

Westfall, P. H., & Young, S. S. (1993). Resampling-based multiple testing. Examples and methods for p-value adjustment.

White, H. (1982). Maximum likelihood estimation of misspecified models. Econometrica, 50(1), 1–25.

Wiens, J. J. (1998). Does Adding Characters with Missing Data Increase or Decrease Phylogenetic Accuracy? Systematic Biology, 47(4), 625–640. https://doi.org/10.1080/106351598260635

Wilks, S. S. (1938) The large sample distribution of the likelihood ratio for testing composite hypotheses. Ann. Math. Stat. 9, 60-62.

Wright, S. (1932). The roles of mutation, inbreeding, crossbreeding and selection in evolution. Proceedings of the Sixth International Congress on Genetics. https://doi.org/citeulike-article-id:1586133

Yang, Z. (2006). Computational molecular evolution. Oxford Series in Ecology and Evolution, XVI, 357.