An Examination of the Transfer of Errors to Species Tree Estimation Caused by Model Selection in Gene Tree Estimation

by

Nevada Basdeo

(Under the Direction of Liang Liu)

Abstract

Inferences from phylogenetic trees is useful in forensic science, bioinformatics, identifying pathogens, and other applications. Thus, building accurate trees is important. Research on nucleotides substitution models has shown the models to be robust for estimating gene trees, but the effects on estimating species trees has not been examined. Cumulative errors on gene tree estimation can transfer over to species tree estimation. Even if the errors are small on each estimated gene tree, they can add up and have a significant impact on accuracy of species tree estimation. In part one of this research, simulations were used to explore how wrongly specified models affect species tree estimation. In part two, data from Austrian finches were used to explore the error of estimation in 30 genes. We found that the models we used in the simulations were robust in species tree estimation. In the finch data, 24 of the 30 estimated genes had a significant chi-square, meaning the 24 genes did not fit the data well. Genes with high GC content appear to have large residuals. Almost all of the residuals were positive suggesting that the evolutionary models were underestimating the frequency of most patterns. Having a vast majority of the genes not being correctly modeled, leads to the adage 'garbage in, garbage out,' in reference to building a species tree. For improvements,

models should better address genes with high GC content and address the under-fitting issue. Due to computational constraints, the results of the simulations may have been affected by the sample size of genes. The simulations might need a bigger sample size of genes to detect an error in species tree estimation if a true error existed.

INDEX WORDS: Phylogenetic Trees, Nucleotide Substitution Models

An Examination of the Transfer of Errors to Species Tree Estimation Caused by Model Selection in Gene Tree Estimation

by

Nevada Basdeo

B.S., Barry University, 2008

A Thesis Submitted to the Graduate Faculty
of The University of Georgia in Partial Fulfillment
of the
Requirements for the Degree

Master of Science

Athens, Georgia

2017

An Examination of the Transfer of Errors to Species Tree Estimation Caused by Model Selection in Gene Tree Estimation

by

Nevada Basdeo

Approved:

Major Professor: Liang Liu

Committee: Cheolwoo Park
T.N. Sriram

Electronic Version Approved:

Suzanne Barbour
Dean of the Graduate School
The University of Georgia
May 2017

An Examination of the Transfer of Errors to Species Tree Estimation Caused by Model Selection in Gene Tree Estimation

Nevada Basdeo

April 24, 2017

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Phylogenetics is the study of evolutionary relationships and history among biological entities, such as species or genes. Typically, phylogenetics examines one of the following questions: (1) what are the evolutionary relationships or histories among species/genes, (2) how do sequences of DNA, RNA, or protein evolve, (3) can the processes of sequence evolution be described better with a mathematical model? Phylogenetics expands our knowledge of genes, genomes, and species evolution. Not only do we learn how the sequence came to be, but we also discover principles that allow us to predict how the sequences will evolve in the future. Applications of phylogenetics include species and genes classification, identifying pathogens, forensic science, and bioinformatics.

Evolutionary relationships can be visualized in phylogenetic trees. There are two types of phylogenetic trees, gene trees and species trees. Gene trees symbolize the evolutionary history of the genes; they can also provide evidence for gene duplication events, as well as evidence for speciation events. Gene trees group alleles of a single gene into phylogeny.

Species trees, which are based on gene trees, depict the ancestral relationships between individuals.

The construction of a phylogenetic tree starts from sequences of different species that are believed to share a common link in their evolutionary history. Before building the phylogenetic trees, the sequences must be aligned. Sequence alignment is a way of arranging the sequences of DNA, RNA, or protein to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences (Yxuhehybyja, 2010). When two symbolic representations of DNA or protein sequences are arranged next to one another, we can identify point mutations of insertion and deletion. If an insertion or deletion occurred in one of the sequences, this would offset the sequence from the rest and can usually be noticed through different lengths of sequences.

Once aligned, there are various approaches for building trees; this research uses the maximum likelihood (ML) method. In the ML method, a heuristic algorithm is used to construct trees of evolutionary history from the observed data, and we calculate the trees' respective probabilities. The tree with the highest probability is identified as the most likely phylogeny. The ML method can require a tremendous amount of computing power making it a slow process, especially for large data sets.

In the ML method, a nucleotide substitution model is chosen when building the gene trees from the sequence data. A nucleotide substitution is a point mutation where a single nucleotide is substituted for a different nucleotide during translation. There are numerous nucleotide substitution models to simulate, or predict, these point mutations. The Markov chain is a stochastic process for modeling nucleotide substitutions. In a Markov chain, the value of the model only depends on the current value and is independent of previous values. Markov models of DNA sequence evolution are used to describe the rate at which one nucleotide replaces another in the evolutionary process. One family of evolutionary models is called the General Time Reversible models (GTR). In the family of GTR models, the

probability of going from event A to event B is the same as going from event B to event A, i.e. $P(i->j) = P(j->i)$, where $i = \{A, C, G, T\}, j = \{A, C, G, T\}$. This family consists of many nested models. We are concerned with three common DNA substitution models in this family: the Jukes and Cantor 1969 model (JC), the Hasegawa, Kishino, and Yano 1985 model (HKY), and the Generalized Time-Reversible model. The JC and HKY models are nested within the GTR model.

The complexity of the model depends on the number of parameters in the model. The parameters for the DNA substitution models are the nucleotide substitution rates and the base frequencies. Nucleotide substation rates refer to the number of nucleotide substitutions per site per unit time. Base frequencies refer to the frequency of adenine (A), cytosine (C), guanine (G), and thymine (T) in the nucleotide sequences. The base frequencies must add up to one; therefore, if three base frequencies are known the fourth is fixed.

The JC is the simplest DNA substitution model, whereas the GTR is the most complex model. The JC model assumes equal mutation rates and equal base frequencies; therefore, there is only one parameter, which represents the nucleotide substitution rate. Unlike the JC model, the HKY model allows for unequal base frequencies. The HKY model has five parameters because it has two nucleotide substitution rates, and three parameters for the base frequencies. In the HKY model the two substitution rates are for transition and transversion rates. The transition rate is the rate of substitution of one purine for another purine or one pyrimidine for another pyrimidine. A transversion is a substitution from a purine to a pyrimidine or vice versa. The GTR is the most complex model with nine parameters, six for the different nucleotide substitution rates and three parameters for the base frequencies. If the more complicated model is not a significantly better fit for the real data, the simpler model is preferred.

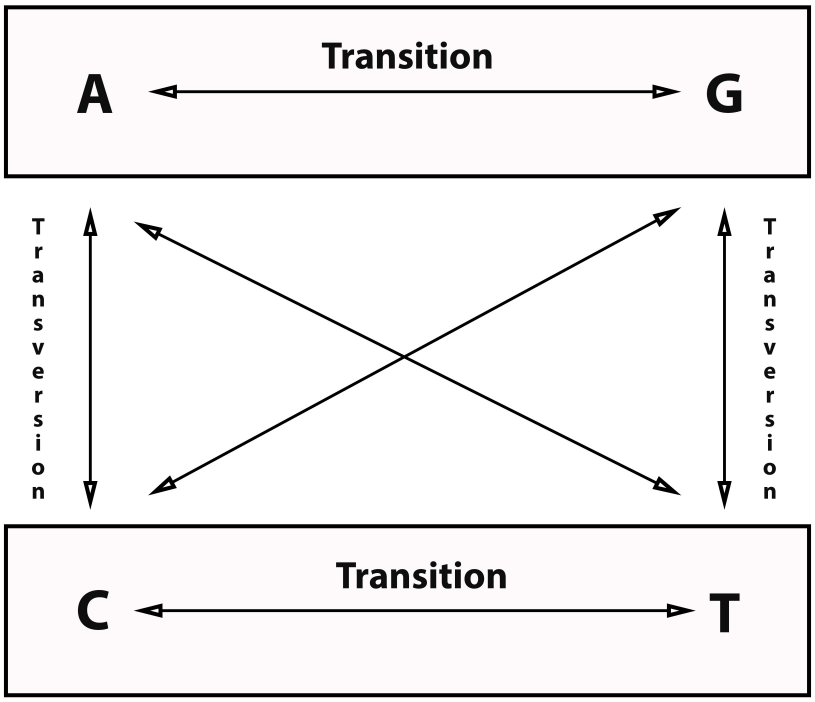


Figure 1.1: Transitions and Transversions

Previous research has focused on model selection for each gene independently (Liu et al., 2008) and has found that model selection does not matter for estimating a single gene tree (Felsenstein, 1981). However, it is currently unknown if model selection matters when estimating a species tree. There has been little attention given to the effect of the chosen substitution model on species tree estimation. In this thesis, the chosen model's goodness-of-fit on the genes is evaluated. A chi-square test is used to compare the multiple sequence alignment patterns (MSAP) expected frequencies to MSAP observed frequencies. A statistically significant difference between the observed MSAP and expected MSAP means that

proposed models with the given parameters are unlikely to produce a distribution similar to that of the observable data. In order to do the test, nucleotides are assumed to be independent of all other nucleotides; the nucleotide is not influenced by nucleotides around it, or at any other site in the sequence.

Chapter 2

Methods

2.1 Simulations

The purpose of these simulations was to examine the effect of using the wrong nucleotide substitution model on species tree estimation. To determine the effects, we started from a specified species tree (the true species tree), from which we generated gene trees (true gene trees), from which we generated nucleotide sequences. From the nucleotide sequences, we generated estimated gene trees, from which we generated the estimated species tree (Figures 2.1 and 2.1). To derive the gene trees from the true species tree, we used the formula developed by Rannala and Yang (2003), which is applied by the function sim.coaltree.sp in the R package phybase. Seq-Gen simulated the evolution of nucleotide sequences. The program read in gene trees and produced nucleotide sequences for each gene tree based on a nucleotide substitution model.

Once the nucleotide sequence was produced, we reversed the procedural order by starting from the nucleotide sequence and ending at the estimated species tree. RAxML, a program for maximum likelihood-based inference of large phylogenetic trees, analyzed the sequences generated from Seq-Gen and produced estimated gene trees, under one of the nucleotide

substitution models. MP-EST built the estimated species tree from a set of estimated gene trees, by maximizing a pseudo-likelihood function. We then compare the estimated species tree to the true species tree. Success is defined when the estimated species tree matches the true species tree.

For the simulations, we use three different modes, JC, HKY, and GTR. When deriving the nucleotide sequences, we chose one of the three models, which was defined as the true model. When estimating the gene trees, we again chose one of the three models, defined as the proposed model. Therefore, there are nine different combinations (JC-JC, JC-HKY, JC-GTR, HKY-JC, etc.) used during the simulations. Examining these nine different combinations in various simulations help us to understand the effects of choosing a wrong model, by under-fitting or over-fitting, on estimating a species tree. We believe that if the proposed model is the same as the true model, the estimated species trees should be closer to the true species tree than estimated species trees created under a different model.

The five simulations used the unrooted species tree ((((A: 0.01, B: 0.01): 0.01, C: 0.02): 0.01, D: 0.03) :0.01, E: 0.04). The length of each generated sequence is 1000 nucleotides. Methods for each of the five simulations are described below.

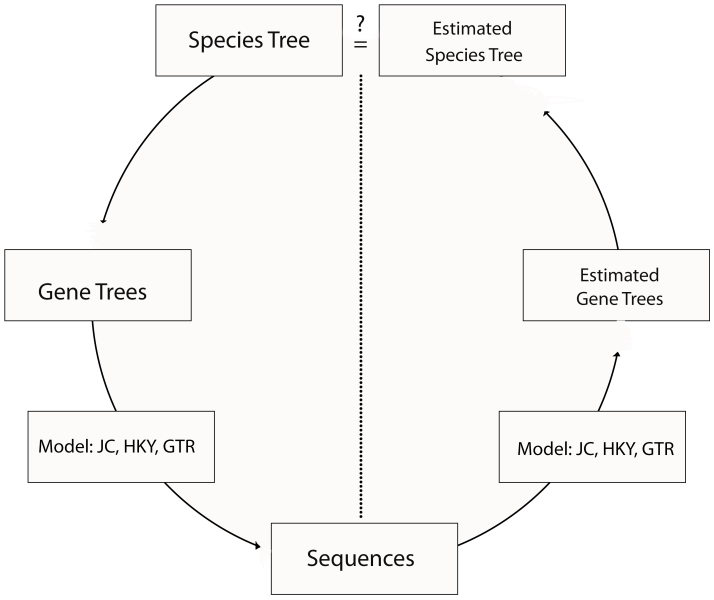


Figure 2.1: Flowchart of Simulation

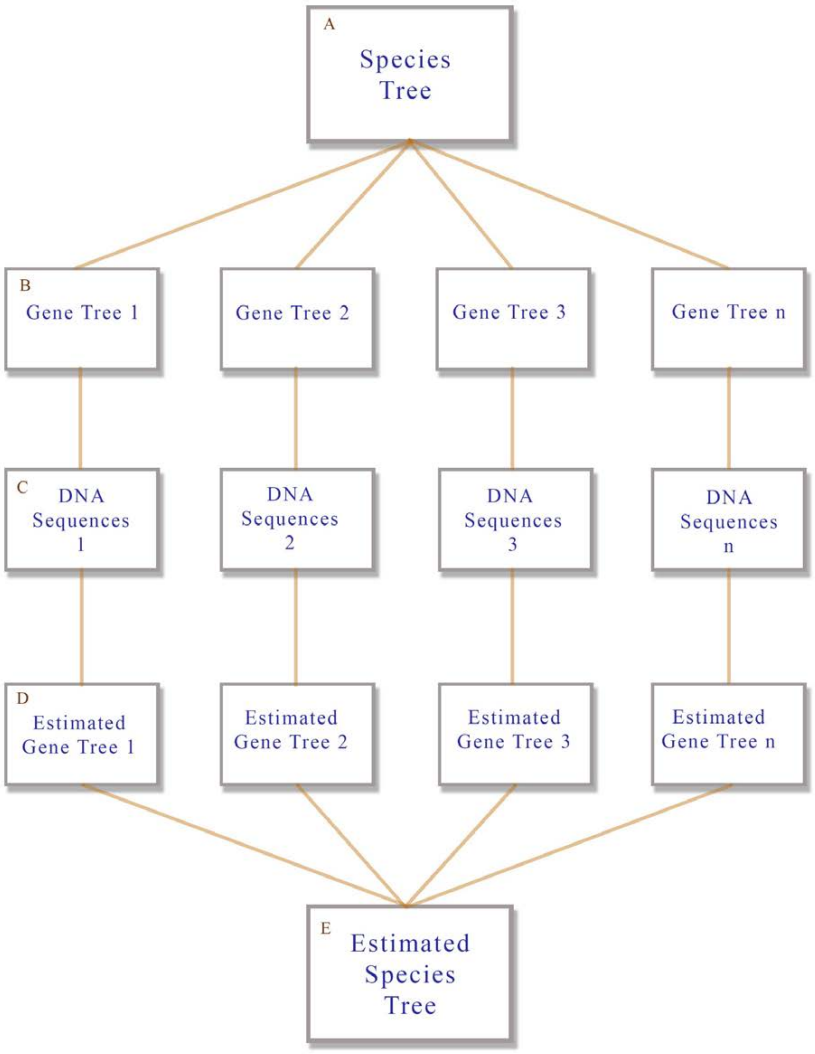


Figure 2.2: Sim Map

2.2 Sim 1

The first simulation is a baseline simulation. This simulation is constrained by keeping many variables constant (non-randomized). The variables that are held constant are mutation rates, branch lengths, nucleotide frequencies, transition and transversion ratio, the shape parameter, and the rate matrix values. Other simulations, which relax one or more of the previously mentioned variables, are compared to this simulation.

2.3 Sim 1A

We simulated 100 data sets each generating 100 true gene trees, giving a total of 10,000 true gene trees in all. Seq-Gen then generated the sequences for the 100 data sets, with RAxML and MP-EST estimating the gene trees and species tree respectively. We compared the 100 estimated species trees to the true species tree, and the 10, 000 estimated gene trees to their respective true gene tree. The base proportions for the models are listed in Table 2.1. The transition-transversion ratio for the JC and HKY model was 0.05 and 0.1 respectively. The GTR rate matrix is listed in table 2.2.

Table 2.1: Nucleotide Base Proportions for Models

Model	A	C	G	T
JC	0.25	0.25	0.25	0.25
HKY	0.08	0.10	0.77	0.05
GTR	0.31	0.22	0.26	0.20

Table 2.2: Relative Rate of Substitution

Base	A	C	G	T
A	0.79			
C	0.11	0.76		
G	0.09	0.05	0.76	
T	0.01	0.08	0.10	0.81

2.4 Sim 1B

In Sim 1B, we used the previously simulated data; but instead, we formed 10 groups each with 1,000 genes. The groups were submitted to MP-EST to estimate the species trees. This regrouping allowed us to explore the effects of a bigger sample size.

2.5 Sim 2

In Simulation 2, instead of using fixed parameters we generate the parameters for the GTR and HKY models from a probability distribution instead of being fixed. For the HKY model, the base frequencies for the nucleotides were randomized from a Dirichlet distribution with its shape parameter equal to 5. For our nucleotie substitution model we randomized our transition - transversion ratio and our shape parameter using a normal distribution with a means of 0.9 and 0.5, respectively and a standard deviation of 0.2 and 0.1, respectively. For the GTR model, the frequencies were randomized from a Dirichlet distribution with its shape parameter equal to 5. We randomized the rate matrix values from a log-normal distribution with a mean of 1.0 and a standard deviation of 1.0. The shape parameter is randomized in the same way as in the HKY model. The JC model for this simulation was omitted, since its parameters are always fixed and remained the same as in Simulation 1. We performed 100 simulations, each with 100 gene trees, resulting in 100 estimated species trees.

2.6 Sim 3

Simulation 3 is the same as Simulation 2, but instead, we double the sample size from 10,000 gene trees to 20,000 gene trees, and we also run the JC model. Simulation 3 consist of two parts, A and B. In part A we run 100 simulations each generating 200 gene trees. In part B we group 2,000 estimated gene trees, forming ten groups. In part A we submit 100 files to MP-EST to get 100 estimated species trees, and in part B, we submit 10 files to MP-EST to get 10 estimated species trees.

2.7 Sim 4

In simulation 4, we allowed variable mutation rates on the different branches of the tree. Variable mutation rates allow us to relax the assumption that the molecular evolution is approximately constant over time for all lineages. We randomized the different branch mutation rates using a log-normal distribution, with a mean of 1 and a variance of one 0.5. We ran 100 simulations, each with 100 gene trees, resulting in 100 estimated species trees.

2.8 Sim 5

In Simulation 5, we examined how shorter branch length affects estimation. We repeated the variable mutation rates across the branches as in Simulation 4, but species branches C and D came from a different log-normal distribution resulting in branches considerably shorter than the rest. For species branch C and D, the log-normal distribution had a mean of 0.1 with a variance of 0.5. We wanted to examine how having two species with shorter branch length affect the estimation. Here we did 100 simulations each with 100 gene trees. We then estimated a total of 10,000 gene trees to predict the species tree.

2.9 Real Data

Australian birds are often studied by biogeographers. In Australia there are many geological barriers causing geographical separation. These geographical separations are believed to be the primary mechanism responsible for speciation as discussed by Keast (1958). In northern Australia there are two barriers, the Carpentarian Barrier and Kimberley Plateau-Arnhem Land Barrier, that may have played a key role in bird diversification. Several closely related species of Australian grass finches in the genus *Poephila* illustrate both of the classic northern Australian biogeographic patterns (Keast, 1958). Jennings and Edwards (2005) collected allelic data obtained from one individual per population of *Peophial acuticauda*, *P. hecki*, and *P. cincta*. They also included sequences for a distant relative, the zebra finch (*P. guttata*). Part of the four Finches DNA sequences were used in this analysis.

The sequences were already aligned, and there was no missing data or gaps. The file was in nexus format, which also included the gene matrix. The Finch data contained 30 genes, and each gene was separated into its own file for analysis. For each gene, we did a frequency count of the MASP's, which was our observable frequency. We submitted each gene file to the Jmodel Test (Darriba et al. (2012) and Guindon and Gascuel (2003)) that uses AIC as its model selection criterion. The best model was chosen out of JC, HKY, and GTR. Under the best model, Phyml (Guindon et al. (2010)) calculated the probabilities of each pattern, and we multiplied this by the sequence length to get the expected frequency counts. A goodness-of-fit chi-square test was done on each gene to see if the model fit the data.

Chapter 3

Statistical Methods

3.1 Theoretical Probabilities

To calculating the probability of an MSAP, we assumed that the columns in the sequences are independent of one another. For k species, there are 4^k different MSAP's. In the finch data, we have four species, thus having 256 different MSAP's. The distribution of the columns follow a multinomial distribution (Eq. 3.1). For a given tree where the branch lengths, terminals, and internodes are known, such as Gene Tree 1 (Figure A.1), we can calculate the MSAP by multiplying, for each path, P_{ij} times the branch length, t_n, starting from the root and leading to a terminal end. Equation A.1 shows the calculation for Gene Tree 1. If the internodes are unknown, as in Gene Tree 2 (Figure A.2), then we must account for all the possibilities at each internode. Each internode would have four possibilities {A, C, G, T}. In Gene Tree 2, there are three unknown internodes, therefore 64 (4^3) different trees that can lead to an 'ACCA' MSAP. Summing up those 64 tree probabilities would give the probability of occurrence of an 'ACCA' column. Note that P_{ij} is used in the above calculations; the next paragraph discusses how the P_{ij}'s are obtained.

$$p(y_1, y_2, ..., y_k) = \frac{n!}{y_1!y_2!..., y_k!} p_1^{y_1} p_2^{y_2} ... p_k^{y_k} \tag{3.1}$$

$$where \sum_{i=1}^{k} p_i = 1 \ \ and \ \ \sum_{i=1}^{k} y_i = n.$$

In calculating the probability of an MSAP, P_{ij} is needed. P_{ij} is the probability of going from one nucleotide to another and can be found in the probability transition matrix, P(t). The probability transition matrix is a 4x4 matrix, where the rows are equal to {A,C,G,T} and the columns are equal to {A,C,G,T}. Each row-by-column element in the P(t) matrix represents a different P_{ij}. The rows of the matrix are considered the inputs and the columns are considered the outputs. However, since we assume a time-reversible process, the matrix is symmetrical with $P_{ij} = P_{ji}$. To find the $P(t)$ matrix we take the exponential of the Q matrix (eq. 3.2).

$$P_{ij}(t) = exp(Qt), \tag{3.2}$$

$$where \ i = \{A, C, G, T\}$$

$$j = \{A, C, G, T\}.$$

The rate transition matrix (Q matrix), is also a 4x4 matrix, with the rows and columns defined as in the $P(t)$ matrix. The elements of the matrix are the rates of moving from one nucleotide to another. Again, the rows are the inputs, and the columns are the outputs.

Each row sums to zeros, where the element in the main diagonal equals the negative sum of the other three elements in the row. To get the Q matrix, a model must be chosen, and then the free parameters of that model are estimated from the data. That is, given a data set, we use the maximum likelihood approach in estimating our parameters for the theoretical model. Once we have our estimated parameters, we can derive our Q matrix. The next paragraphs discuss the JC, HKY and GTR model and how to obtain their respective Q matrices.

$$P_{ij}(t) = \begin{cases} \frac{1}{4} + \frac{3}{4}exp(-\mu t) & \text{if } i = j \\ \frac{1}{4} - \frac{1}{4}exp(-\mu t) & \text{if } i \neq j. \end{cases} \tag{3.3}$$

The key variables in nucleotide substitution models are the substitution rate(s) from one nucleotide to another nucleotide and the time frame over which substitution could occur. Since straightforward time points are not usually available for molecular data, and we cannot separate the time and rate variables, the product of rate and time, known as genetic distance, is more commonly used. We reparametrize the genetic distance as τ. Distinctions between the models include how they regulate both base frequencies and substitution rates.

$$Q_{JC} = \begin{pmatrix} -\frac{3}{4}\mu & \frac{1}{4}\mu & \frac{1}{4}\mu & \frac{1}{4}\mu \\ \frac{1}{4}\mu & -\frac{3}{4}\mu & \frac{1}{4}\mu & \frac{1}{4}\mu \\ \frac{1}{4}\mu & \frac{1}{4}\mu & -\frac{3}{4}\mu & \frac{1}{4}\mu \\ \frac{1}{4}\mu & \frac{1}{4}\mu & \frac{1}{4}\mu & -\frac{3}{4}\mu \end{pmatrix}. \tag{3.4}$$

Model JC has two distinct assumptions. Frist, we assume that nucleotide substitutions are equally likely, meaning that the substitution rates are the same between all nucleotides. Second, we assume that base frequencies are equal among the four nucleotides (25%). As a result of these assumptions, the JC model has one parameter, μ, denoting the overall substitution rate. The JC Q matrix is shown in equation (3.4) and the elements of the $P(t)$ matrix is shown in equation 3.3.

The HKY model incorporates multiple parameters to create a more realistic model of how nucleotide sequences evolve by relaxing the two assumptions discussed in the JC model: we allow transitions and transversions to occur at different rates, and we allow base frequencies to vary relative to each other. The Q matrix for the HKY model is shown in matrix 3.5. In the HKY Q matrix, the k parameter indicates a transversion substitution. In most sequence comparisons, transitions are found to occur more frequently than transversions, even though each nucleotide has two transversions and only one transition. Transitions may be more common since cells have mechanisms to detect mismatched base pairs. Four of the nucleotide substitution equations are shown in A.2, where v is the branch length in terms of the expected number of changers per site. The other state combinations can be obtained by substituting in the appropriate base frequencies.

$$Q_{HKY} = \begin{pmatrix} * & \pi_C & k\pi_G & \pi_T \\ \pi_A & * & \pi_G & k\pi_T \\ k\pi_A & \pi_C & * & \pi_T \\ \pi_A & k\pi_C & \pi_G & * \end{pmatrix}. \tag{3.5}$$

The GTR model is the most complex model of the group. Like the HKY model, the nucleotides can occur at different frequencies. However, in the GTR model, each pair of nucleotide substitutions occurs at a different rate. Since the model is time reversible, A changes to T at the same rate that T changes into A. Since the model is time reversible, the probability transition matrix is symmetrical.

$$Q_{GTR} = \begin{pmatrix} * & \delta\pi_C & \eta\pi_G & \beta\pi_T \\ \delta\pi_A & * & \epsilon\pi_G & \alpha\pi_T \\ \eta\pi_A & \epsilon\pi_C & * & \gamma\pi_T \\ \beta\pi_A & \alpha\pi_C & \gamma\pi_G & * \end{pmatrix} \tag{3.6}$$

where

$$\alpha = r(T \to C) = r(C \to T)$$

$$\beta = r(T \to A) = r(A \to T)$$

$$\gamma = r(T \to G) = r(G \to T)$$

$$\delta = r(C \to A) = r(A \to C)$$

$$\epsilon = r(C \to G) = r(G \to C)$$

$$\eta = r(A \to G) = r(G \to A).$$

3.2 Two-Proportion Z-Interval

For the simulations, we used hypothesis testing to make statistical inference about our population proportions. The point estimate of a population proportion, p, was given by $\hat{p} = \frac{x}{n}$, where x is the number of correct estimated species trees, and n is the total number of esti-

mated species trees in the simulation. The simulations were considered to be independent samples. A two-proportion z-interval (Eq. 3.7) was used to determine if the proportions of correctly estimated species tree were statistically different between true-model and estimated model pairings. If the confidence interval contained zero for the difference between two model pairings, then the proportions were not considered to be statistically different. A 95% confidence interval was used for the calculations. We also assumed the sampling distribution of $\hat{p}$ is approximately normal.

$$\hat{p_1} - \hat{p_2} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}} \tag{3.7}$$

3.3 AIC

The Akaike information criterion (AIC) provides a means for model selection and is based on the Kullback-Leibler divergence (Posada and Buckley, 2004). AIC estimates the information lost when a particular model is used to represent the process that generates the data. As such, it deals with the trade-off between the complexity of the model and the goodness of fit of the model. AIC compares the relative quality of nested statistical models for a given data set. Given a collection of nested models, AIC estimates the relative quality of each model to all other candidate models. The model with the lowest AIC score is considered the best model *relative* to the other tested models. AIC does not measure the *absolute* quality of the model; if all the candidate models fit poorly, AIC will simply choose the best among the group of poorly fitting models.

$$AIC = 2(k) - 2ln(\hat{L}) \tag{3.8}$$

We used AIC to evaluate gene tree models of the Finch data. Models JC, HKY, and GTR were compared, and the model with lowest AIC score for a particular gene was used to estimate that gene tree.

3.4 Goodness-of-Fit Chi-Square Test

To examine how closely our proposed model fits the finch data, we performed a GOF chi-square test. We performed the analysis on all thirty genes independently. For each gene, we compared the observed counts of MSAP's to our expected counts of MSAP's. To get our expected counts of MSAP's we multiplied the probability of an MSAP by the sequence length. A statistically significant difference between the observed frequencies and the expected frequencies would suggest that our model data differed from the observed finch data. The degrees of freedom for the chi-square test was n-1, where n was the number of patterns. In the observed data, some MSAP's were not observed, and all of these unobserved patterns were grouped together in one bin called 'REST.' For instance, if our observed data had ten different patterns, our n would be 11 for the ten patterns plus one for the REST. The probability of the REST was the sum of all the unobserved MSAP's probabilities. The chi-square test should have an expected count of five in each bin, but this was not the case in the finch data. Some of the patterns had such a small probability coupled with a short to moderate sequence length that resulted in the expected counts being below five or even zero. Consequently, we also conducted a conditional chi-squared test, conditioned on the observable patterns. a GOF χ^2 test (Eq. 3.9). We used the Bonferroni correction since we were simultaneously performing the GOF test on the finch data. This correction adjusted the p-value to 0.05/30.

$$\chi^2 = \sum_{all\,cells} \frac{(Obs - Exp)^2}{Exp} \tag{3.9}$$

Chapter 4

Analysis and Results

4.1 Sim 1a

The results suggest that over-fitting nor under-fitting have an adverse effect. Even though the proportions of over-fitting tend to be slightly higher, there was not a statistical difference between the proportions; the differences could be due to chance. Table 4.1 list the proportions of correctly estimated species trees from simulation 1a. The column headers list the true model, where as the rows list the estimated model. To further investigate if any underlying patterns existed, we increase the sample size of gene trees used to estimate the species tree in Simulation 1b, with the proportion results listed in Table 4.2 . Table B.1 list the proportion of estimated gene trees that matched the true gene tree for all the simulations. The results in Table B.1 are in line with previous research that suggests that the models are robust for gene tree estimation. Table B.2 shows the average topological distance and standard deviation of estimated gene tree from the true gene tree. Table B.2 shows the average topological distance and standard deviation of estimated gene tree from the true gene tree. Estimated gene trees were close to their respective gene trees.

Table 4.1: Sim 1a Proportion of Correctly Estimated Species Tree

Model	1.JC	1.HKY	1.GTR
2.JC	.55	.53	.56
2.HKY	.55	.40	.61
2.GTR	.59	.44	.67

4.2 Sim 1b

The results from increasing the number of gene trees used to estimate the species tree may show hints of some issues with under-fitting. However, the sample size is now too small to determine if the proportions are truly different from one another or the difference is just due to chance. When GTR is the true model, under-fitting with the JC model may cause problems and additional simulations would be needed to determine if there is a statistically significant difference.

Table 4.2: Sim 1b

model	1.JC	1.HKY	1.GTR
2.JC	.70	.40	.30
2.HKY	.50	.40	.50
2.GTR	.70	.30	.60

4.3 Sim 2

Randomizing the parameters for the HKY and GTR models, the GTR model start to show some patterns of being more accurate Table 4.3. When GTR is the true model, under-fitted models appear to perform worse at predicting the species tree (p-value= 0.058 for the difference between the proportions between HKY and GTR). This finding would be in line with our expectations that under-fitting models would lead to more errors in predicting the true species tree. Increasing the simulations in order to increase the sample size may give a more definitive conclusion.

Table 4.3: Sim 2

model	1.HKY	1.GTR
2.JC	.59	.58
2.HKY	.58	.55
2.GTR	.65	.69

Table 4.4: Sim 2 Correctly Estimated Gene Trees

model	1.HKY	1.GTR
2.JC	.18	.16
2.HKY	.18	.17
2.GTR	.18	.18

4.4 Sim 3a

In this sample, we increased the sample size of estimated gene trees to examine the effects on estimating species trees. The results (Table 4.5) again show there is no statistical difference between the models in estimating species tree. Under-fitting or over-fitting does not seem to be of any concern. When HKY is the true model and the JC model is used (under-fitting) there may be some issues of concern. Increasing the number of simulations could shed light on the issue.

Table 4.5: Sim 3a

model	1.JC	1.HKY	1.GTR
2.JC	.81	.74	.74
2.HKY	.79	.84	.78
2.GTR	.82	.80	.74

4.5 Sim 3b

Table 4.6: Sim 3b

model	1.JC	1.HKY	1.GTR
2.JC	.7	.7	.8
2.HKY	.5	.8	.9
2.GTR	.6	.8	.7

The results (Table 4.6) are in line with the previous findings in that there is no statistical difference between the different models estimating the correct species tree. Previous research has shown that using the wrong model does not adversely affect estimating the gene trees; and thus results in this thesis support this claim when estimating species trees.

4.6 Sim 4

Even when mutation rates were allowed to vary in the model, there was no statistical difference in under-fitting or over-fitting models. This results (Table 4.7) is consistent with results of prior simulations.

Table 4.7: Sim 4

model	1.JC	1.HKY	1.GTR
2.JC	.62	.54	.50
2.HKY	.62	.55	.51
2.GTR	.61	.48	.58

4.7 Sim 5

In this simulation, we shorten the branch length of two species, making the lengths unequal. Overall these models did worse than the equal-length models in predicting the correct species tree (Table 4.8). Models HKY and GTR seem to be more affected by these unequal branch lengths than the JC model. Again, there was no statistical difference between the models in estimating the correct species tree.

Table 4.8: Sim 5

model	1.JC	1.HKY	1.GTR
2.JC	.50	.51	.56
2.HKY	.50	.43	.50
2.GTR	.43	.46	.44

4.8 Finch Descriptive Statistics

In the Finch data, as previously mentioned there were no gaps or missing data, and the file contained 30 genes for the finches. The sequence length for each species was 16,119 nucleotide bases. A chi-square test for homogeneity revealed that there was no statistical difference between the frequencies of nucleotides among the different species (x=0.106, df = 9, p-value = 1 see Table 4.9). There were more adenines and thymines than cytosines and guanines in each species (see Figure 4.1). The most common MSAP's were TTTT, AAAA, CCCC, and GGGG. MASP's with three of the same nucleotides were the next most common patterns.

4.9 AIC

The Jmodel Test used AIC as a means of model selection. Table 4.11 shows the results for the 30 genes. Model HKY was chosen as the best model for nineteen of the genes, JC for six

genes, and the GTR model for five genes. Results of the model selection were treated as the theoretical model and submitted to phyml to get the expected probabilities of all MSAP's.

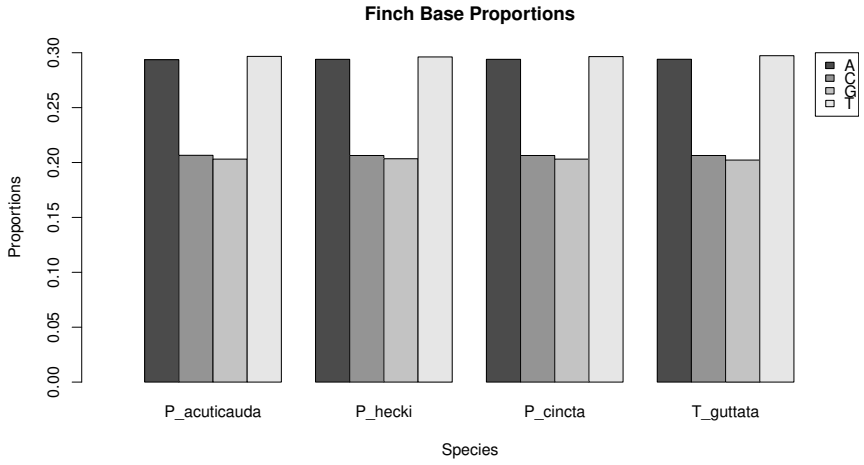


Figure 4.1: Nucleotide Base Proportions for the Finches

Table 4.9: Species' Nucleotide Base Frequencies

Base	P_acuticauda	P_Hecki	P_cincta	T_guttata
A	4733	4739	4739	4740
C	3330	3327	3327	3327
G	3274	3279	3274	3260
T	4782	4774	4779	4792

4.10 Goodness of Fit

Our goodness-of-fit results show that our estimated genes do not accurately reflect the observed data. For our regular chi-square test ($\alpha = 0.05$), 24 of the 30 estimated genes were statistically different from the observed genes, even after the Bonferroni correction (Table 4.12). For our conditional chi-square test ($\alpha = 0.05$), 21 of the estimated genes were significantly different, and after the Bonferroni correction, 16 were still significantly different from

Table 4.10: Frequencies of Most Common Patterns

MSAP	Count
TTTT	4652
AAAA	4615
CCCC	3208
GGGG	3137
GGGA	51
CCCT	42
AAAG	40
TTTC	32

Table 4.11: AIC Model Selection for Finch Genes

Best Model	Genes	Count
JC	5, 6, 7, 14, 22, 23	6
HKY	1, 3, 4, 9 ,11, 12, 13, 15, 16, 17, 18, 19, 20, 21, 25, 26, 27, 28, 30	19
GTR	2, 8, 10, 24, 29	5

the observed genes. These differences supports the notion that even after model selection a GOF test should be performed. With most of the estimated gene trees not being accurate, these errors are likely to transfer over to estimating species trees. Table 4.12 shows the p-values for the chi-square tests. The '*' in the table indicates a significantly different chi-square test under the Bonferroni correction.

Table C.1 shows the MSAP that had standardized residuals that were above or below 4 standard deviations of the expected counts of MSAP's . The 'REST' term in the table refers to all of the MSAP's that were not observed in the real data. When calculating the chi-square, we summed all the probabilities of the non-observed patterns and called it the the probability of observing the rest. From Table C.1 we notice the MSAP's with GC content tend to have the most extreme residuals. Appendix C.1 includes the standardized residual plots for each gene of the finch data. The residuals are not symmetrical around the y=0 line.

Most of the residuals lie above the line, meaning that our proposed models are consistently underestimating many frequencies of the MSAP's.

Table 4.12: GOF Results

Gene	χ^2	Bonferroni Correction	Conditional χ^2	Bonferroni Correction Conditional
1	5.329×10^{-6}		1.590×10^{-04}	
2	3.422×10^{-11}		6.109×10^{-10}	
3	9.973×10^{-27}		3.217×10^{-24}	
4	7.537×10^{-9}		1.978×10^{-07}	
5	1.019×10^{-30}		8.979×10^{-28}	
6	6.680×10^{-6}		8.881×10^{-05}	
7	4.702×10^{-19}		1.034×10^{-17}	
8	0.137*	*	0.485*	*
9	3.900×10^{-11}		1.474×10^{-10}	
10	0.437*	*	0.649*	*
11	0.003		0.019	*
12	2.234×10^{-27}		2.597×10^{-24}	
13	0.087*	*	0.181*	*
14	0.161*	*	0.371*	*
15	8.938×10^{-4}		4.762×10^{-03}	*
16	0.005		0.073*	*
17	8.453×10^{-9}		2.173×10^{-07}	
18	5.662×10^{-56}		2.694×10^{-53}	
19	0.004		0.032	*
20	0.010		0.068*	*
21	3.13×10^{-80}		1.884×10^{-75}	
22	4.328×10^{-4}		3.494×10^{-03}	*
23	0.015		0.102*	*
24	5.639×10^{-67}		1.342×10^{-64}	
25	0.007		0.043	*
26	1.755×10^{-18}		1.746×10^{-17}	
27	0.466*	*	0.437*	*
28	0.059*	*	0.157*	*
29	1.204×10^{-24}		2.273×10^{-20}	
30	8.630×10^{-4}		1.643×10^{-03}	

Chapter 5

Discussion

The quality of an estimator is assessed by its bias and variance. Ideally, we want an estimator to be unbiased with a small variance. To evaluate the goodness of our estimators, we employ the average square of the distance between the estimator and its target parameter, this is known as the mean square error (MSE). The MSE is a function of the bias and variance (Eq. 5.1). The MSE may help to explain why a wrong model may have done just as good or even better than the correct model. The wrong model may have had a lower MSE than the correct model. For instance, the JC may be biased when the true model is the GTR model, but its variance could be smaller than that of the GTR model. The more complicated evolutionary models tend to have a higher variance, due to the greater number of parameters.

$$MSE(\hat{\theta}) = [B(\hat{\theta})]^2 + V(\hat{\theta}) \tag{5.1}$$

Additionally, the simulations results may be unexpected due to sample size limitations. The errors in the estimated gene trees are so small that there might not be enough cumulative errors to affect species tree estimation. In species tree estimation, the error of each gene tree may accumulate, so the more estimated gene trees used, the more errors are transferred over to species tree estimation. The idea is that good estimates of gene trees can lead to a

good estimate of a species tree, however poor estimates of gene trees may result in a poor estimate of a species tree. The sample size in gene trees and repetitive simulations may not be sufficient to detect an effect. This finding suggests that nucleotide substitution models may be robust in estimating species trees when the estimation is drawn from a small number of estimated gene trees. In this study, sample size limitations were due to computing power and computing restrictions. The zcluster has a limit of the number of produced output files. RAxML produces an abundance of output files when estimating gene trees, which may approach the limits of zcluster. Other software should be considered for estimating the gene trees such as Phyml.

The finch data had no sequence gaps. Typically, this would not be the case, as there would be point mutations in the sequences. Calculating MSAP's with gaps would be more difficult. One possible method would be to look at the marginal probabilities for the MSAP's that contain one or more missing nucleotides. For instance, if there are four species and one of the MSAP's is 'AATX,' where X indicates a missing value, then we would count the number of AAT's in the pattern for the first three species. Then we would need to calculate the expected probability of observing this pattern for the first three species, before computing the chi-square. This calculation could be very cumbersome especially when the number of species increases.

With regards to the finch data, when looking at the 24 estimated genes that are rejected, we noticed that GC content might be the cause of poop model fit as patterns with high GC content result in models with the most bias. Patterns with high GC content seems to give us the biggest errors. Table C.1 list the MSAP's with residuals greater than three standard deviations for each gene and Appendix C.1 has a standardized residual plot for each gene. Some of the residuals are as far as 15 to 20 standard deviations away from zero, with quite a few residuals hovering around 8 to 10 standard deviations away from the zero. GC content may have a biological importance as prior research shows that GC content significantly affects evolutionary modeling of nucleotide sequences (Smarda et al., 2014). Models that better handle the GC content would most likely improve the accuracy of estimation.

Even though the first part of this thesis may not have produced statistically significant results, there is value gained from these simulations. Previous research has shown that single gene tree estimation is robust to the nucleotide substitution model, but research has not been done on the robustness of the nucleotide substitution model when estimating species trees. The simulations performed in this research give insight on how choosing the wrong model affects the accuracy of species tree estimation. The findings suggest that species tree estimation is robust to the evolutionary model, similar to the findings of gene tree estimation. However, this finding may not hold up as the sample size of gene trees used to derive the species tree increases. Also, our analysis just considers the topology of the trees. Estimation would be improved taking into consideration the topology and the divergence time. Branch length would likely be affected by the model used, making the errors statistically significant. Future research should consider divergence time and topology in exploring bias in estimating species trees.

Appendix A

Statistical Methods

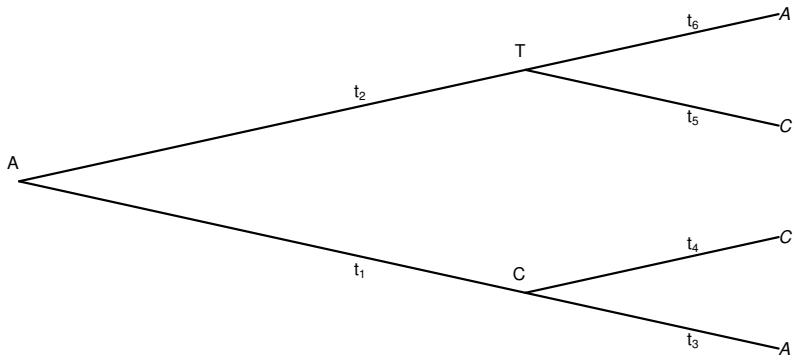


Figure A.1: Gene tree 1 with known internodes

$$P(Gene\ tree\ 1) = [P_{AC}(t_1)P_{CA}(t_3)][P_{AC}(t_1)P_{CC}(t_4)][P_{AT}(t_2)P_{TC}(t_5)][P_{AT}(t_2)P_{TA}(t_6)] \tag{A.1}$$

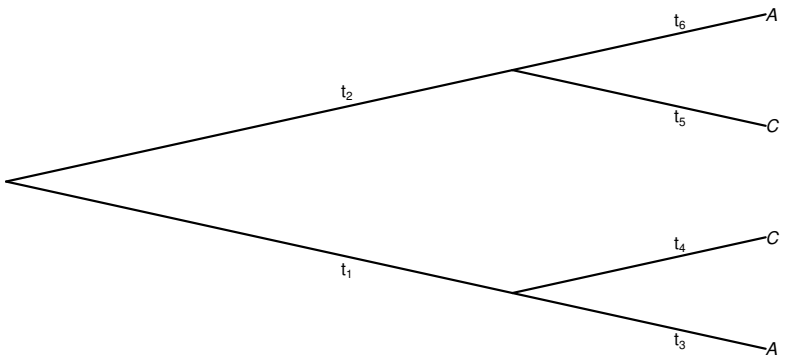


Figure A.2: Gene tree 2 with unknown internodes

$$\beta = \frac{1}{2(\pi_A + \pi_G)(\pi_C + \pi_T) + 2k[(\pi_A\pi_G) + (\pi_C\pi_T)]} \tag{A.2}$$

$$P_{AA}(v,k,\pi) = [\pi_A(\pi_A + \pi_G + (\pi_C + \pi_T)e^{-\beta v}) + \pi_G e^{-(1+(\pi_A+\pi_G)(k-1.0))\beta v}]/(\pi_A + \pi_G)$$

$$P_{AC}(v,k,\pi) = \pi_C(1.0 - e^{-\beta v})$$

$$P_{AG}(v,k,\pi) = [\pi_G(\pi_A + \pi_G + (\pi_C + \pi_T)e^{-\beta v}) - \pi_G e^{-(1+(\pi_A+\pi_G)(k-1.0))\beta v}]/(\pi_A + \pi_G)$$

$$P_{AT}(v,k,\pi) = \pi_T(1.0 - e^{-\beta v})$$

Appendix B

Simulations

B.1 Gene Tree Estimation

Table B.1 shows that proportions of estimated gene trees that matched the true gene trees. The column headers are the true model. Our results are in line with previous research that suggests that the models are robust for gene tree estimation. Table B.2 shows the average topological distance and standard deviation of estimated gene tree from the true gene tree.

Table B.1: Proportion of Correctly Estimated Gene Trees

	Model	JC	HKY	GTR
Simulation 1	JC	.19	.07	.07
	HKY	.19	.07	.07
	GTR	.19	.07	.07
Simulation 2	JC	-	.18	.16
	HKY	-	.18	.17
	GTR	-	.18	.18
Simulation 3	JC	.18	.17	.16
	HKY	.18	.17	.17
	GTR	.18	.17	.17
Simulation 4	JC	.18	.18	.16
	HKY	.19	.18	.16
	GTR	.18	.18	.17
Simulation 5	JC	.17	.16	.15
	HKY	.17	.16	.15
	GTR	.16	.16	.15

Table B.2: Gene Trees Topology Distance

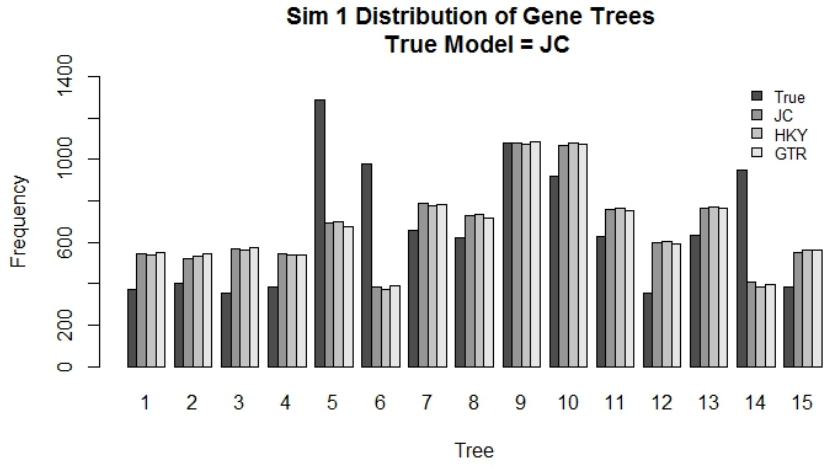
		JC		HKY		GTR	
		Mean	SD	Mean	SD	Mean	SD
Simulation 1	JC	2.44	1.48	3.16	1.25	3.17	1.24
	HKY	2.44	1.49	3.17	1.23	3.17	1.23
	GTR	2.45	1.48	3.17	1.23	3.18	1.23
Simulation 2	JC	-	-	2.49	1.47	2.56	1.46
	HKY	-	-	2.48	1.47	2.54	1.47
	GTR	-	-	2.48	1.47	2.50	1.48
Simulation 3	JC	2.46	1.48	2.51	1.46	2.57	1.44
	HKY	2.47	1.48	2.49	1.46	2.54	1.46
	GTR	2.47	1.48	2.50	1.46	2.52	1.46
Simulation 4	JC	2.46	1.48	2.50	1.47	2.56	1.45
	HKY	2.45	1.48	2.49	1.48	2.55	1.45
	GTR	2.47	1.48	2.49	1.48	2.51	1.46
Simulation 5	JC	2.56	1.46	2.59	1.44	2.67	1.43
	HKY	2.57	1.47	2.57	1.46	2.64	1.44
	GTR	2.57	1.46	2.58	1.45	2.62	1.45

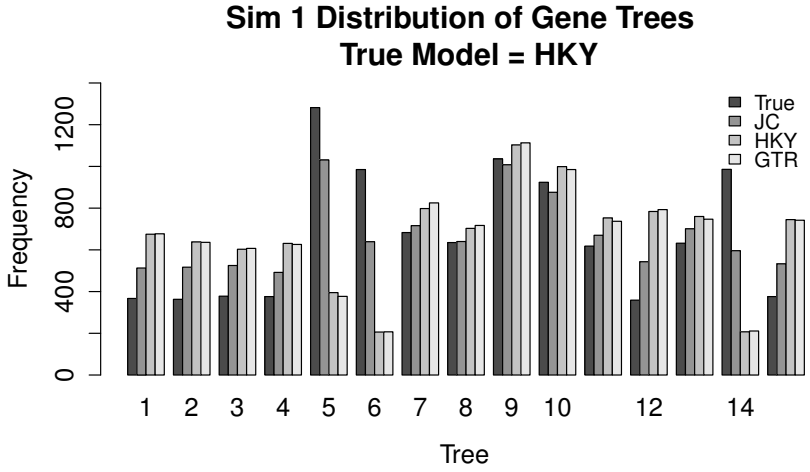
B.2 Possible Trees

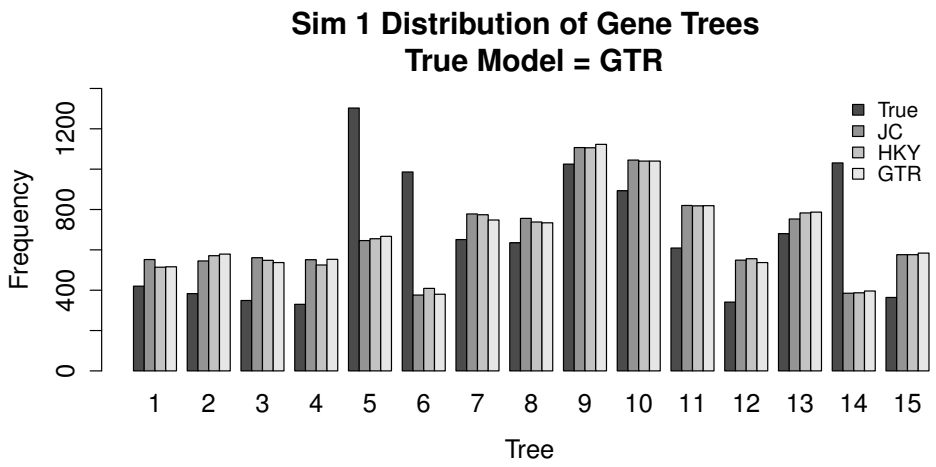
These are the possible trees for simulations. The distributions are graphed according to the order listed below.

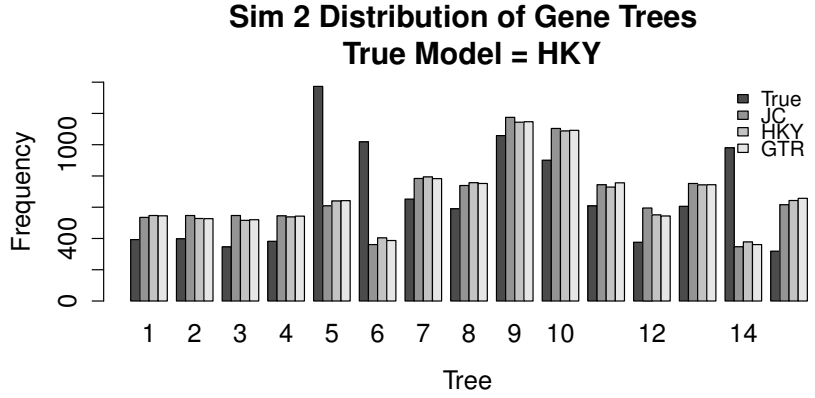
1. ((A,(C,(B,D))),E)
2. (((B,(A,D)),C),E)
3. (((C,(A,D)),B),E)
4. ((((B,D),A),C),E)
5. (((B,A),(D,C)),E)
6. (((A,C),(B,D)),E)
7. ((((C,A),B),D),E)
8. ((A,(D,(C,B))),E)
9. (((C,(B,A)),D),E)
10. (((D,(A,B)),C),E)
11. (((D,(A,C)),B),E)
12. ((A,(B,(C,D))),E)
13. ((((C,B),A),D),E)
14. (((A,D),(B,C)),E)
15. (((A,(D,C)),B),E)

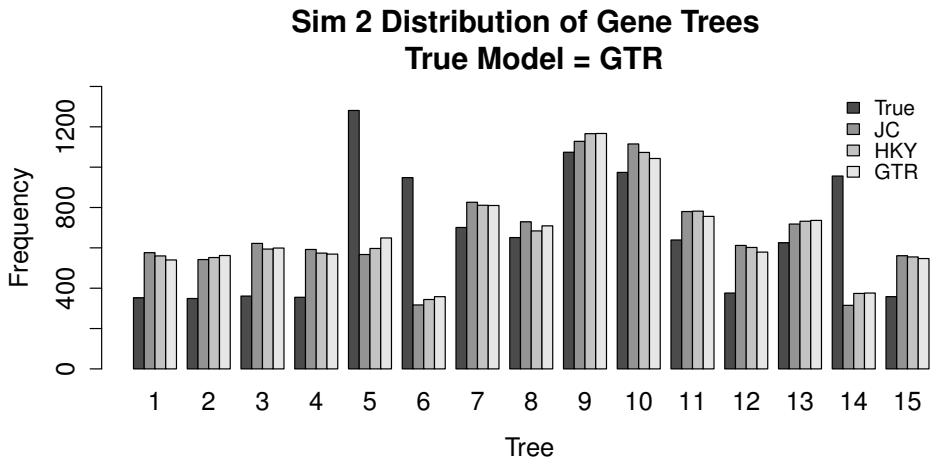
B.3 Distribution of Gene Trees (True & Estimated)

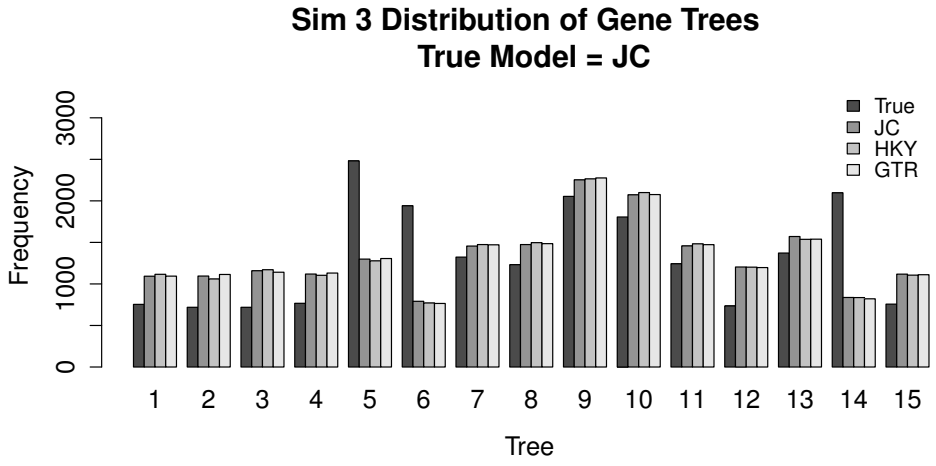


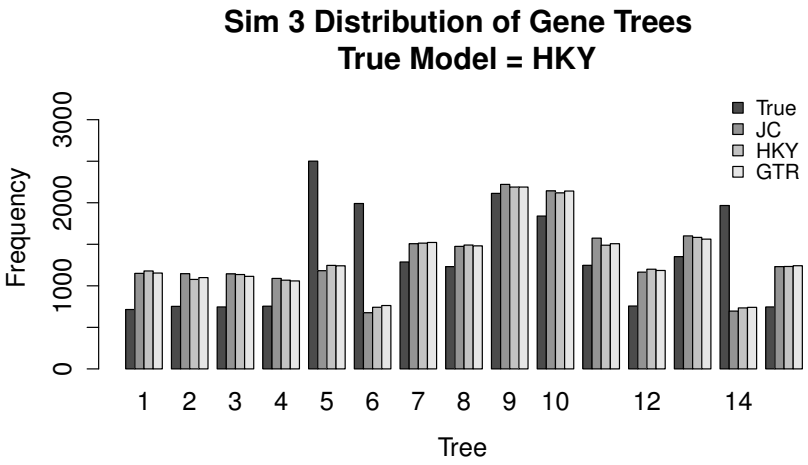


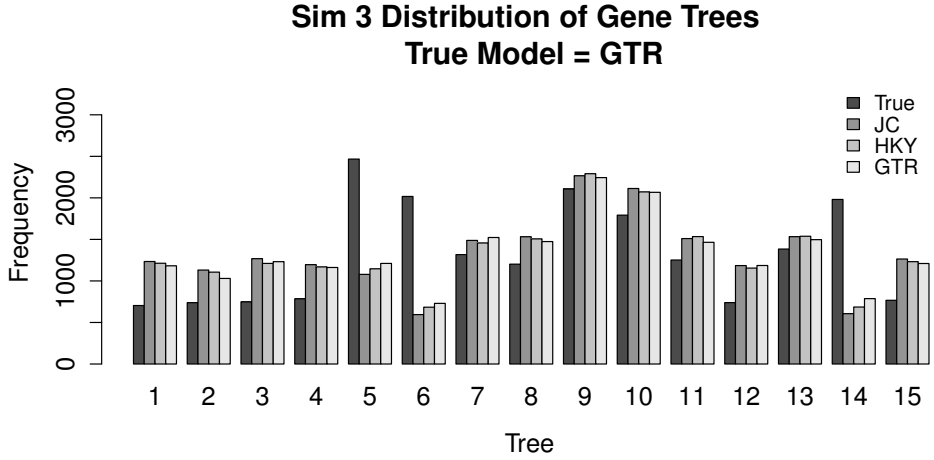

Sim 1 Distribution of Gene Trees
True Model = GTR
Frequency
0
400
800
1200
True
JC
HKY
GTR
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
Tree

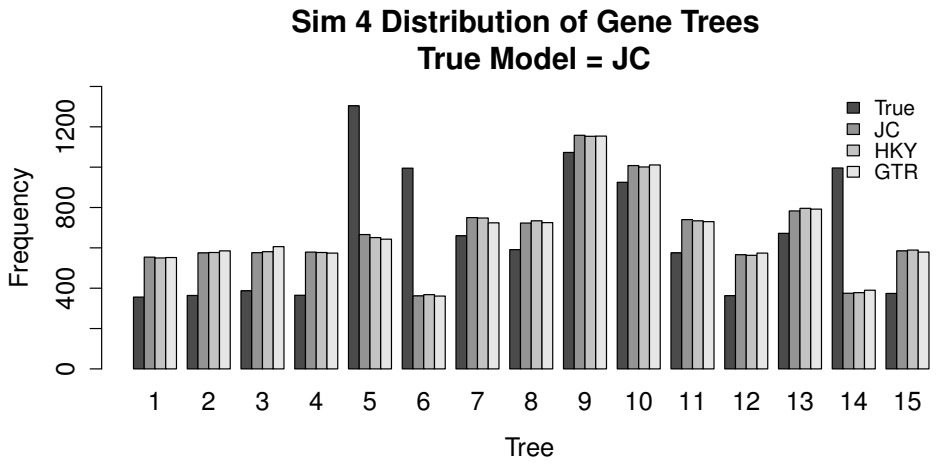

Sim 2 Distribution of Gene Trees
True Model = HKY
Frequency
0
400
1000
True
JC
HKY
GTR
1
2
3
4
5
6
7
8
9
10
12
14
Tree

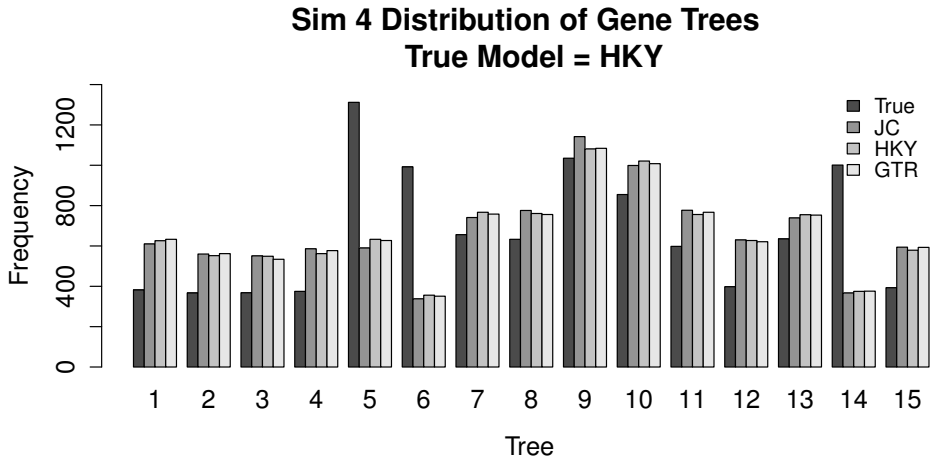

Sim 2 Distribution of Gene Trees
True Model = GTR
Frequency
0
400
800
1200
True
JC
HKY
GTR
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
Tree

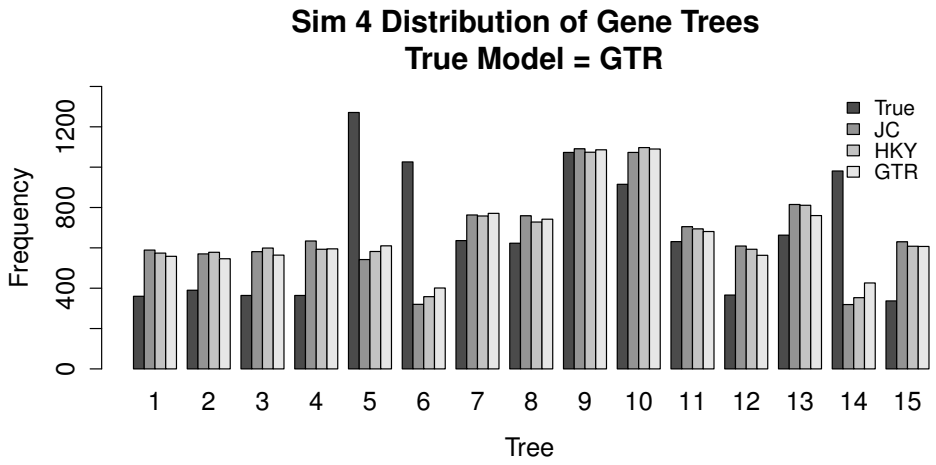

Sim 3 Distribution of Gene Trees
True Model = JC
Frequency
0
1000
2000
3000
True
JC
HKY
GTR
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
Tree

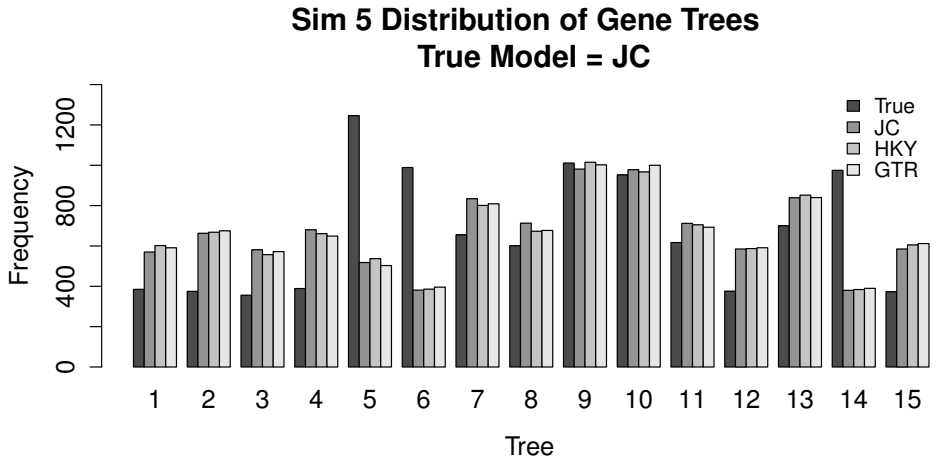

Sim 3 Distribution of Gene Trees
True Model = HKY
True
JC
HKY
GTR
Frequency
0
1000
2000
3000
1
2
3
4
5
6
7
8
9
10
12
14
Tree

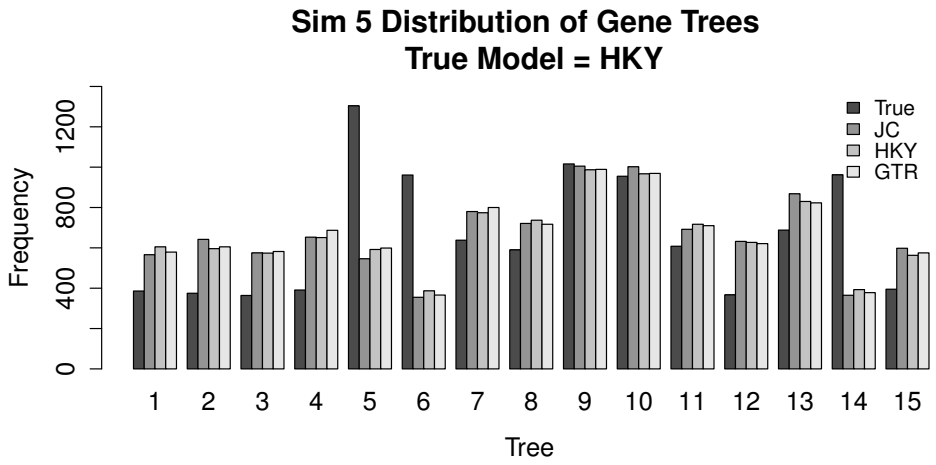

Sim 3 Distribution of Gene Trees
True Model = GTR
True
JC
HKY
GTR
Frequency
0
1000
2000
3000
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
Tree

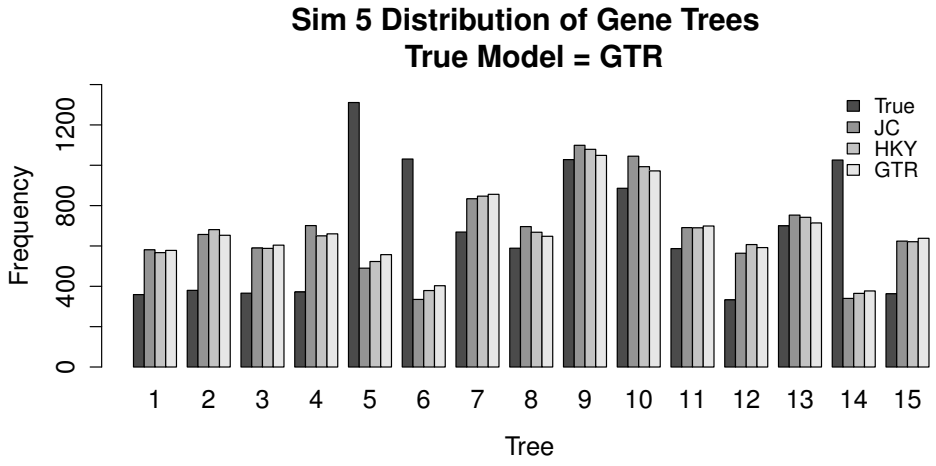

Sim 4 Distribution of Gene Trees
True Model = JC
Frequency
0
400
800
1200
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
Tree
True
JC
HKY
GTR


Sim 4 Distribution of Gene Trees
True Model = HKY
Frequency
0
400
800
1200
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
Tree
True
JC
HKY
GTR


Sim 4 Distribution of Gene Trees
True Model = GTR
Frequency
0
400
800
1200
True
JC
HKY
GTR
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
Tree


Sim 5 Distribution of Gene Trees
True Model = JC
Frequency
0
400
800
1200
True
JC
HKY
GTR
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
Tree


Sim 5 Distribution of Gene Trees
True Model = HKY
Frequency
0
400
800
1200
True
JC
HKY
GTR
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
Tree


Sim 5 Distribution of Gene Trees
True Model = GTR
Frequency
0
400
800
1200
True
JC
HKY
GTR
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
Tree

Appendix C

Goodness-of-Fit

C.1 Multiple Sequence Alignment Patterns and Standardized Residual Plots

Table C.1: Error Prone MSAP's

Finch Gene	Multiple Sequence Alignment Patterns
1	ATTT, CGCC, REST
2	CGCG
3	CGCT, REST
4	ACCC, GGGT
5	GGAC, TTTG
6	CAAA
7	GGAA, GTGT, TTGT
8	
9	CCGC, GTGG
10	
11	
12	CGGT, TCTT, TGGC, REST
13	
14	
15	ACCC, CCGG
16	AAAG, REST
17	TTCA, REST
18	GGCC, TGGT, REST
19	
20	GGGC
21	GGCC, GTGT, REST
22	GCGG
23	
24	GAGG, TCAA, REST
25	AACC
26	CAAA, CTCT
27	
28	
29	CACC, CGGT, GGAT, GGCG, TTCC, REST
30	TAAA

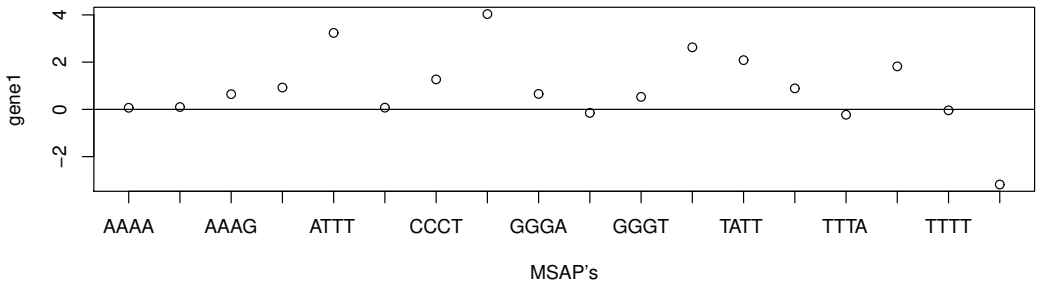


Figure C.1: Gene 1 Residual plot

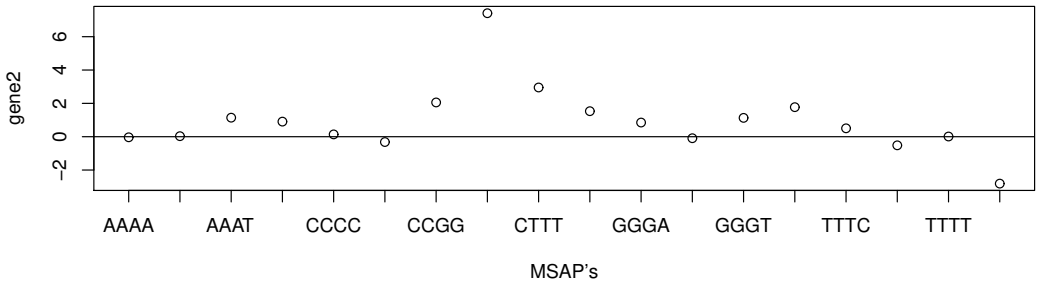


Figure C.2: Gene 2 Residual plot

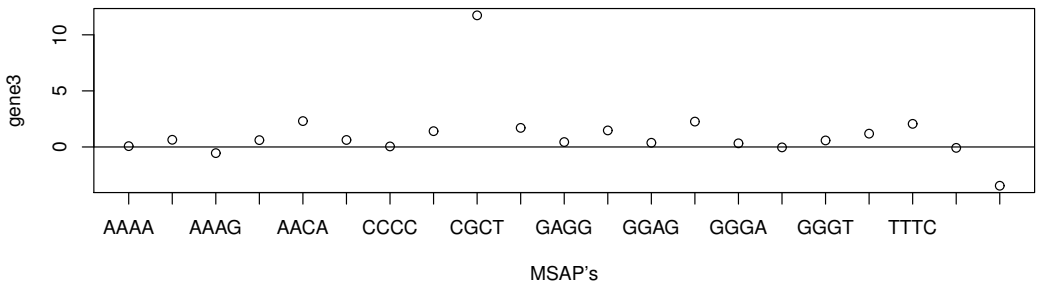


Figure C.3: Gene 3 Residual plot

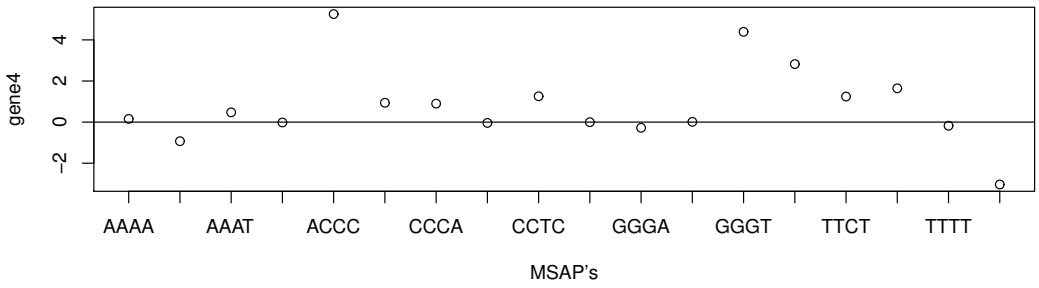


Figure C.4: Gene 4 Residual plot

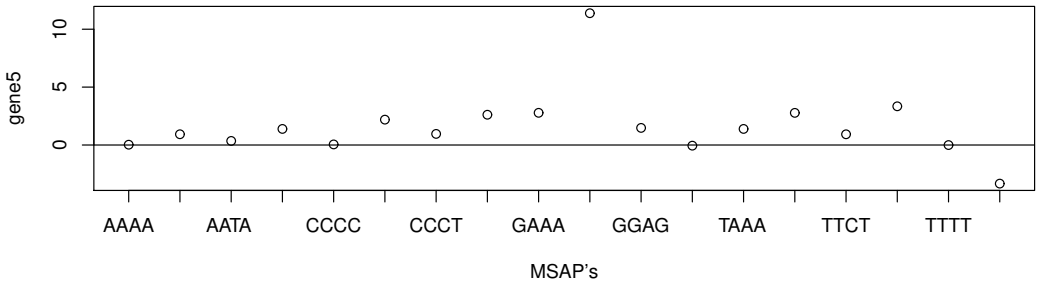


Figure C.5: Gene 5 Residual plot

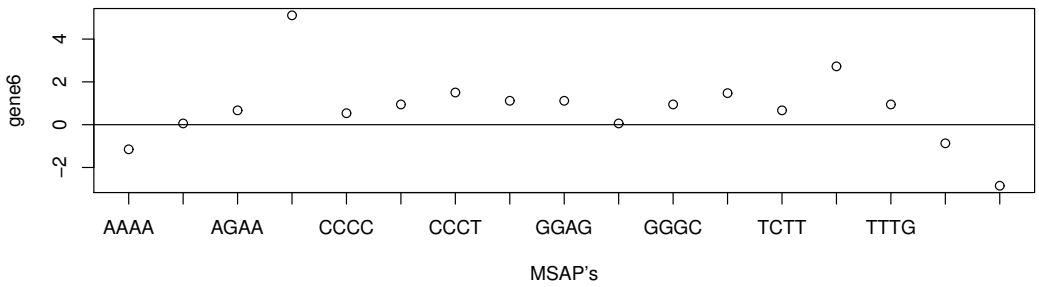


Figure C.6: Gene 6 Residual plot

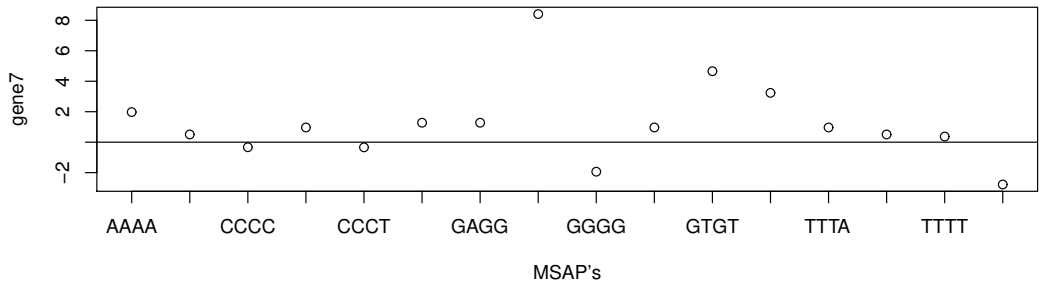


Figure C.7: Gene 7 Residual plot

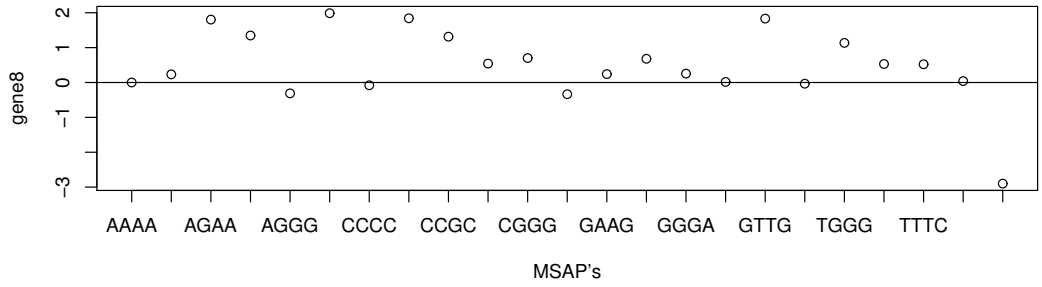


Figure C.8: Gene 8 Residual plot

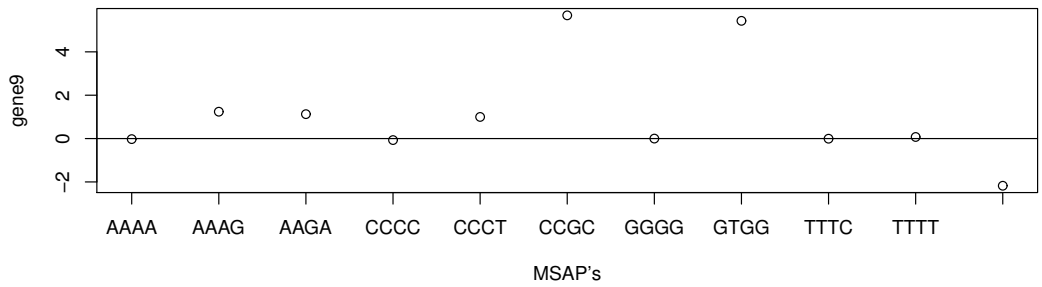


Figure C.9: Gene 9 Residual plot

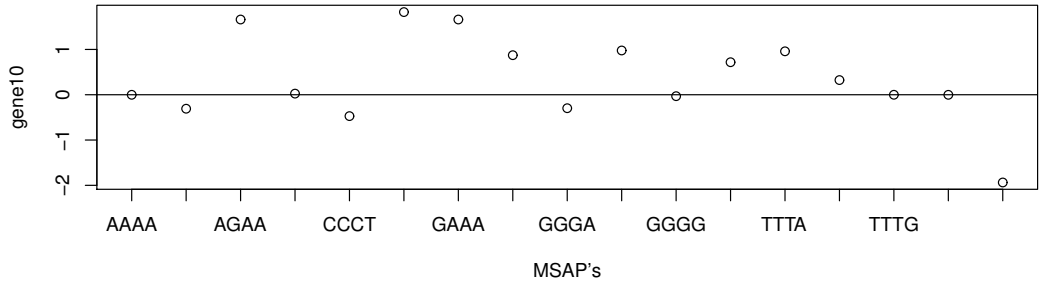


Figure C.10: Gene 10 Residual plot

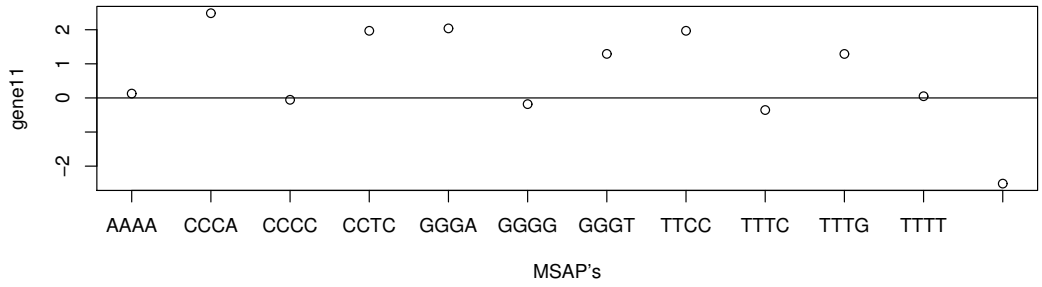


Figure C.11: Gene 11 Residual plot

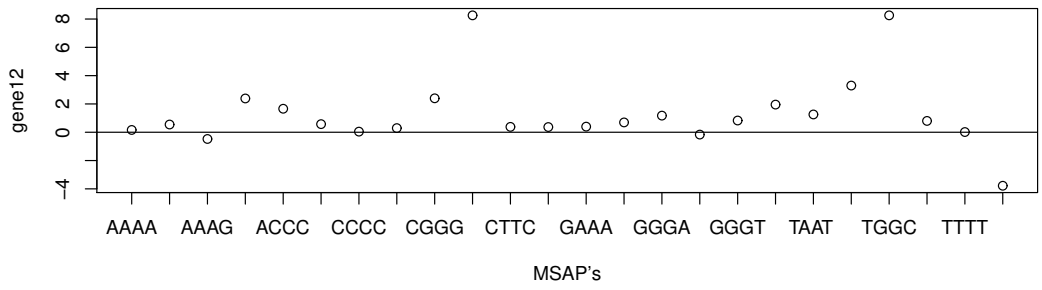


Figure C.12: Gene 12 Residual plot

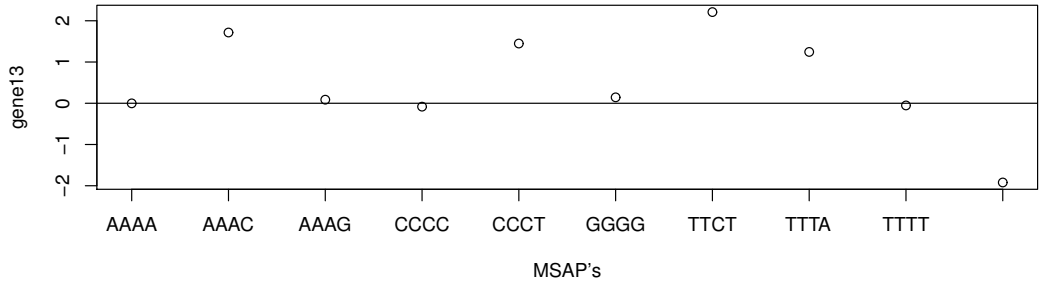


Figure C.13: Gene 13 Residual plot

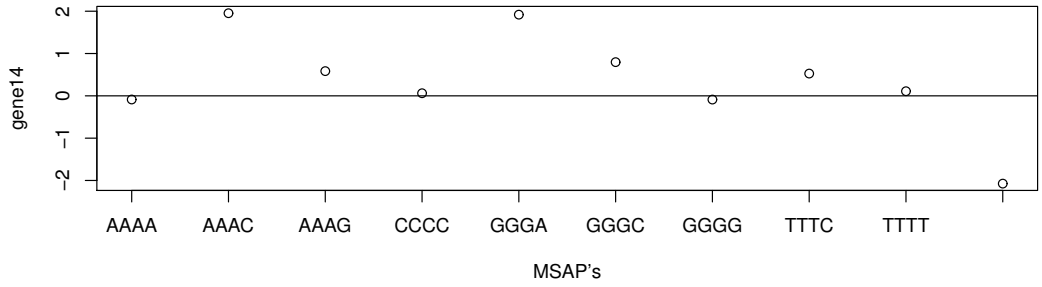


Figure C.14: Gene 14 Residual plot

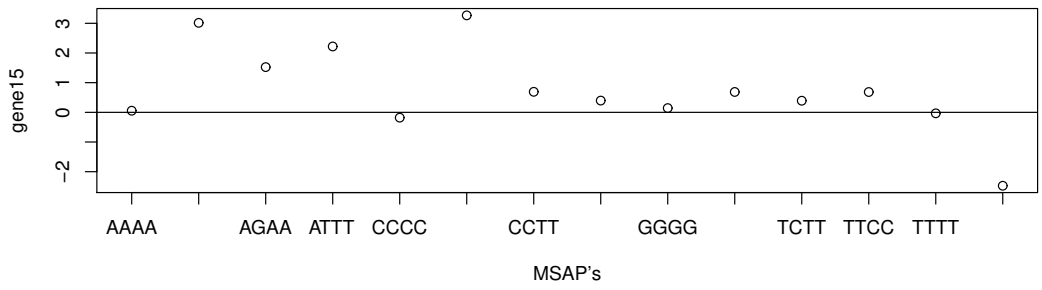


Figure C.15: Gene 15 Residual plot

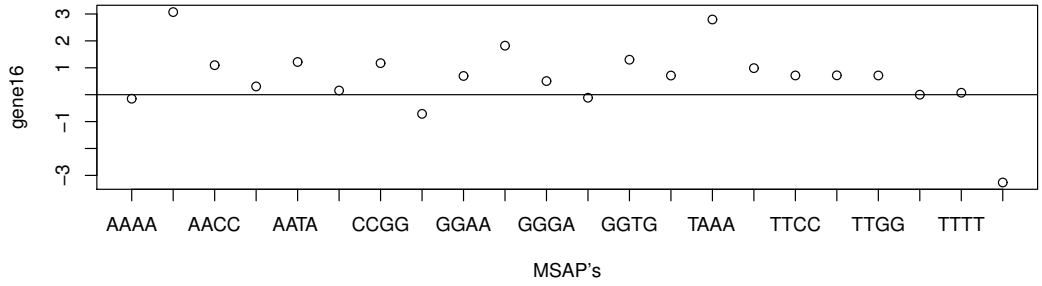


Figure C.16: Gene 16 Residual plot

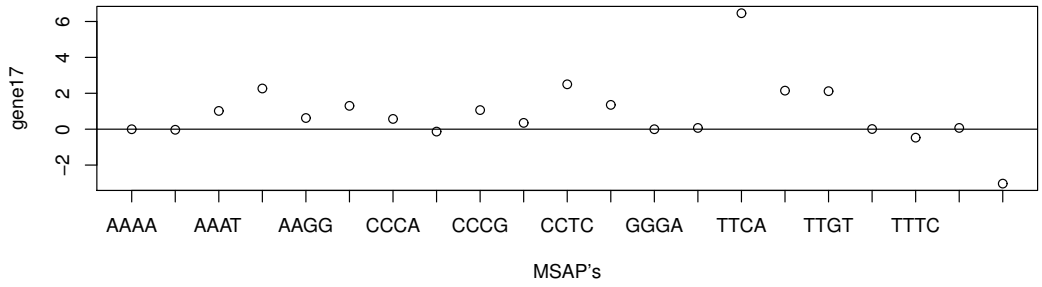


Figure C.17: Gene 17 Residual plot

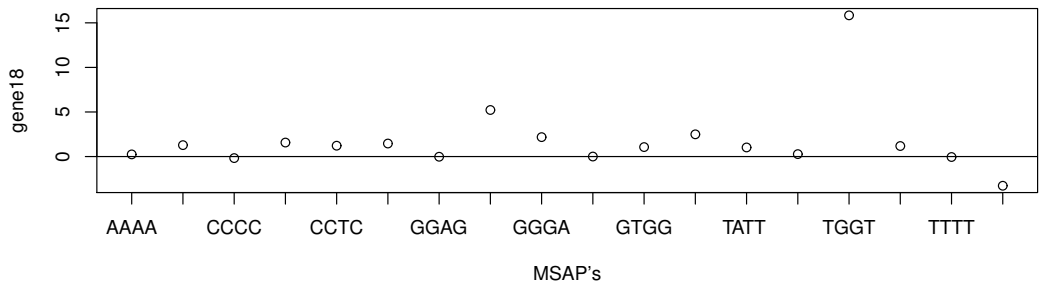


Figure C.18: Gene 18 Residual plot

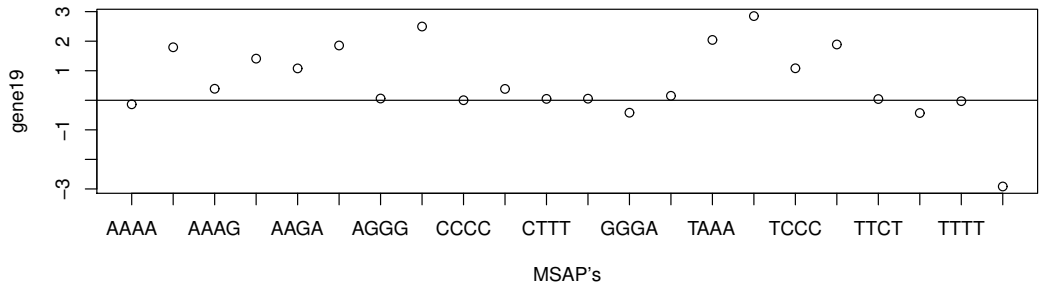


Figure C.19: Gene 19 Residual plot

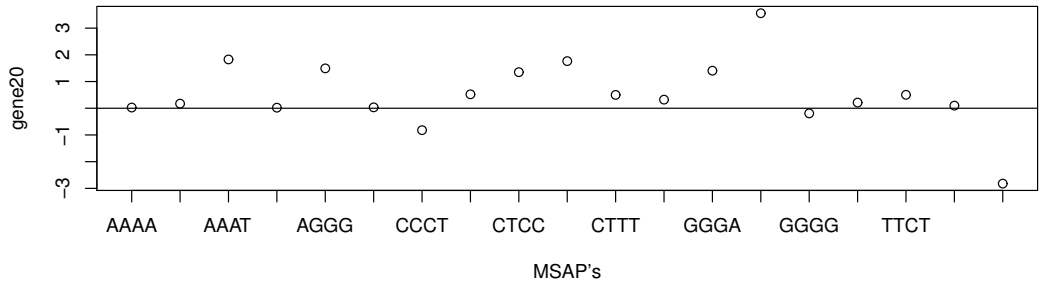


Figure C.20: Gene 20 Residual plot

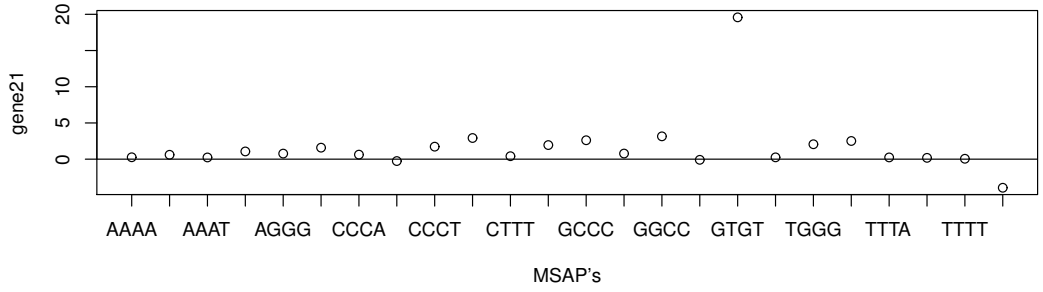


Figure C.21: Gene 21 Residual plot

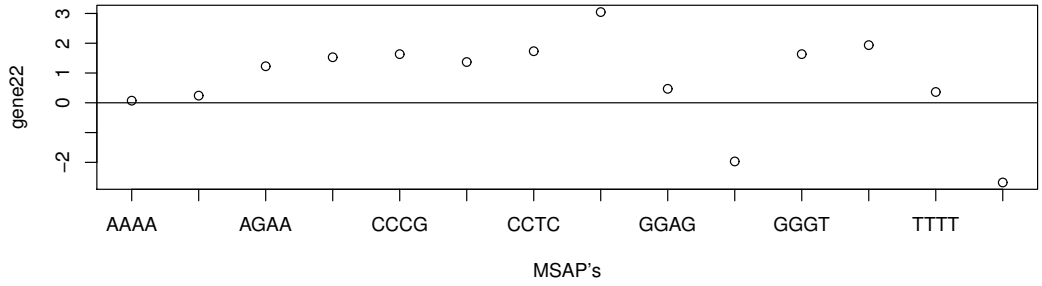


Figure C.22: Gene 22 Residual plot

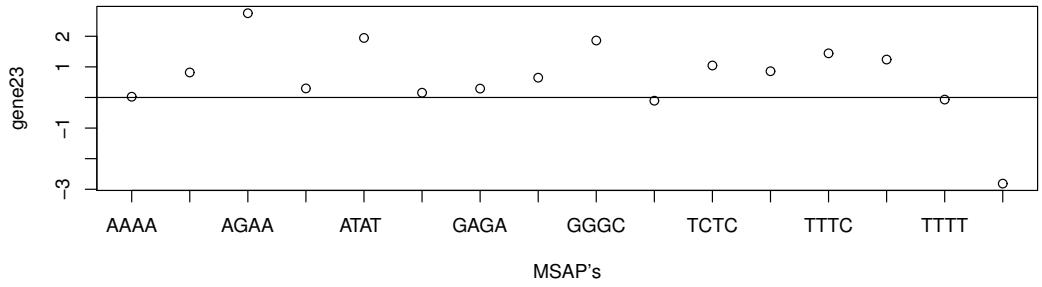


Figure C.23: Gene 23 Residual plot

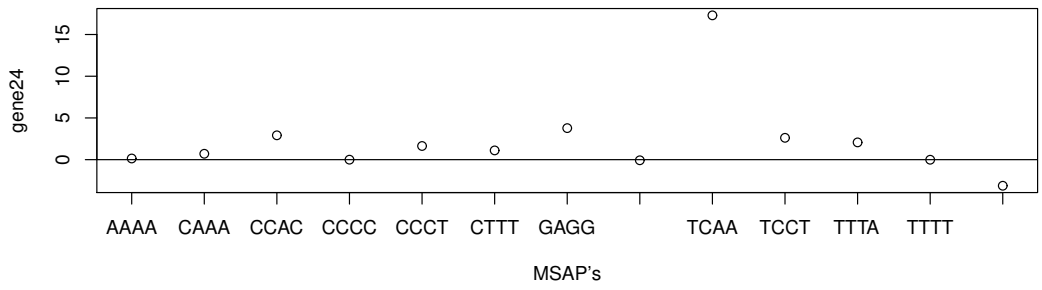


Figure C.24: Gene 24 Residual plot

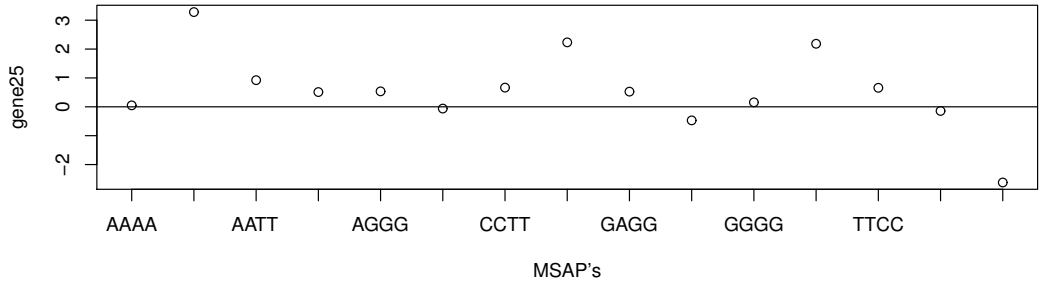


Figure C.25: Gene 25 Residual plot

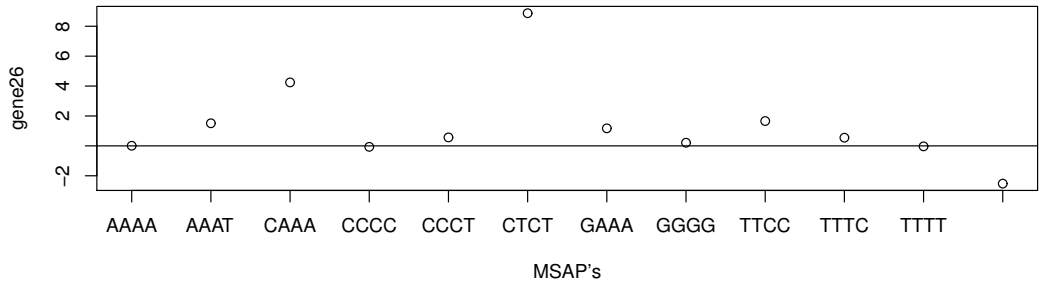


Figure C.26: Gene 26 Residual plot

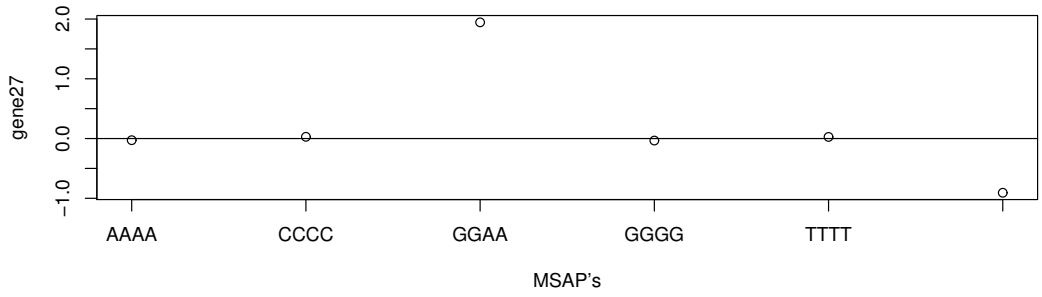


Figure C.27: Gene 27 Residual plot

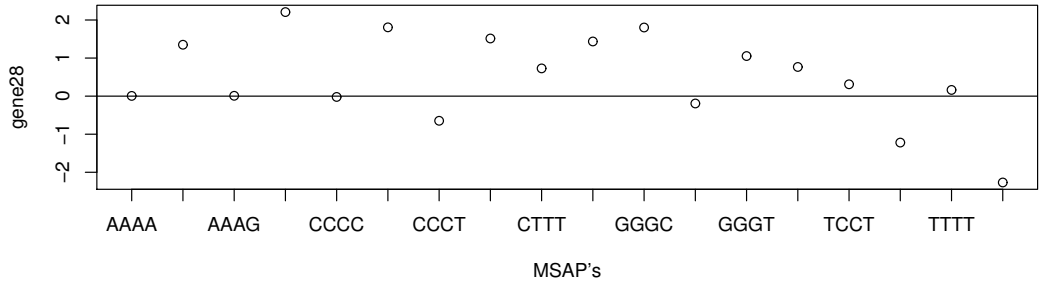


Figure C.28: Gene 28 Residual plot

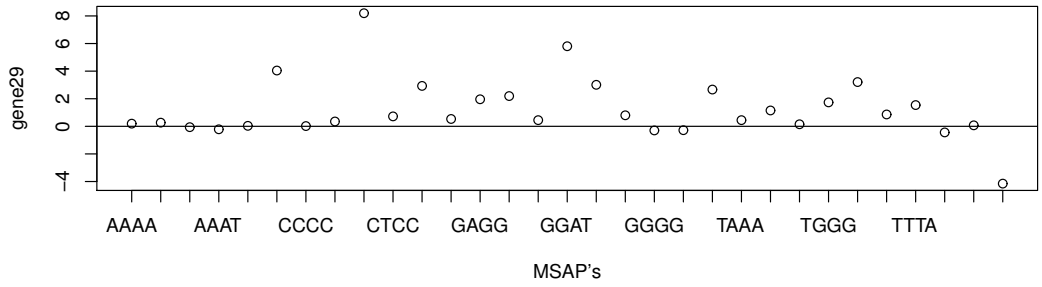


Figure C.29: Gene 29 Residual plot

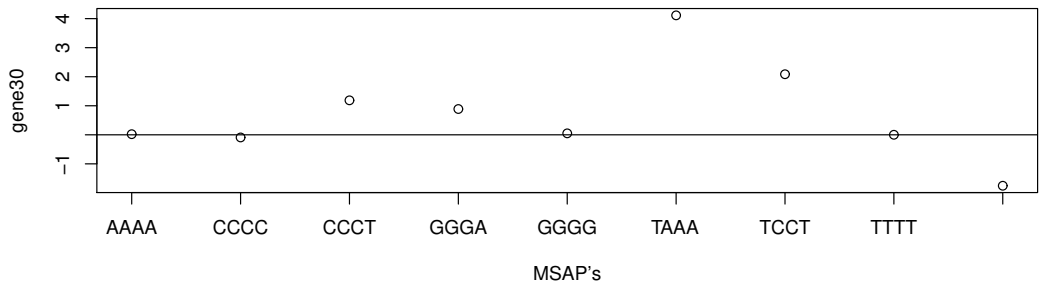


Figure C.30: Gene 30 Residual plot

C.2 Sample of R code

```
library ("phybase") # zcluster
sptree <- "((((A: 0.01, B: 0.01): 0.01, C: 0.02):0.01, D: 0.03):0.01, E:0.04)"
spname <- c("A", "B", "C", "D", "E");
nodematrix <- read.tree.nodes(str=sptree, name=spname)$nodes
nodematrix[,5]<-0.2
nodematrix[8,5]<-0.001
nspecies <- length(spname)
ngenetrees <- 1000
seqlength <- 1000
nsim  <- 1000
for (j in 1:(nsim * ngenetrees)){
  genetrees <- sim.coaltree.sp(rootnode=rootoftree(nodematrix), nodematrix=nodematrix,
                               nspecies=nspecies,seq=rep(1,nspecies), name=spname)$gt
  write(genetrees,paste("truegenetree",j,sep=""))
  freq<-paste(rdirichlet(1,c(5,5,5,5)),collapse=" ")
  shape<-abs(rnorm(1,0.5,0.1))
  tratio<-abs(rnorm(1,0.9,0.2))
  try(system(paste("seq-gen -a", shape, " -g4 -mHKY -f ", freq, " -t", tratio, " -l1000"
}
# produces seq's and truegenetree's
# The four frequencies are generated from the dirichelt distribution (n,a) n is the numb
# Shape parameter and transition/transversion ratio for seq-gen generated from normal di

for(i in 1:nsim)
```

```
{
    raxmlcommand1<-paste("raxml -mGTRGAMMA -s seq",((i-1)*ngenetrees+1):(i*ngenetrees),"
    raxmlcommand2<-paste("raxml -mGTRGAMMA -s seq",((i-1)*ngenetrees+1):(i*ngenetrees),"
    raxmlcommand3<-paste("raxml -mGTRGAMMA -s seq",((i-1)*ngenetrees+1):(i*ngenetrees),"
    write(c(raxmlcommand1,raxmlcommand2,raxmlcommand3),paste("run",i,sep=""))
}
# produce run files
write(paste("qsub -q rcc-30d run",1:nsim,sep=""),"submit")


#-------------------------------------
# 2 root gene trees
#-------------------------------------
genetrees<-1:(nsim*ngenetrees)
for(j in 1:(nsim*ngenetrees)){
    phy<-read.tree(paste("RAxML_bestTree.jc",j,sep=""))
    a<-root(phy, outgroup="A", resolve.root=T)
    a<-root(a, outgroup="E", resolve.root=T)
    a$edge.length <- NULL
    a$node.label <- NULL
    a$root.length <- NULL
    genetrees[j] <- write.tree(a)
}
genetrees<-matrix(genetrees,ngenetrees,nsim)
for(j in 1:nsim) write(genetrees[,j],paste("genetrees_jc",j, sep=""))
# gives genetrees's for JC
```

```
genetrees<-1:(nsim*ngenetrees)
for(j in 1:(nsim*ngenetrees)){
    phy<-read.tree(paste("RAxML_bestTree.hky",j,sep=""))
    a<-root(phy, outgroup="A", resolve.root=T)
    a<-root(a, outgroup="E", resolve.root=T)
    a$edge.length <- NULL
    a$node.label <- NULL
    a$root.length <- NULL
    genetrees[j] <- write.tree(a)
}
genetrees<-matrix(genetrees,ngenetrees,nsim)
for(j in 1:nsim) write(genetrees[,j],paste("genetrees_hky",j, sep=""))
# genetrees for HKY

genetrees<-1:(nsim*ngenetrees)
for(j in 1:(nsim*ngenetrees)){
    phy<-read.tree(paste("RAxML_bestTree.gtr",j,sep=""))
    a<-root(phy, outgroup="A", resolve.root=T)
    a<-root(a, outgroup="E", resolve.root=T)
    a$edge.length <- NULL
    a$node.label <- NULL
    a$root.length <- NULL
    genetrees[j] <- write.tree(a)
}
genetrees<-matrix(genetrees,ngenetrees,nsim)
for(j in 1:nsim) write(genetrees[,j],paste("genetrees_gtr",j, sep=""))
```

```
# gives gene trees for GTR


#-------------------------------------
# 3 ESTIMATING SPECIES TREES USING MPEST
#-------------------------------------
c<-"A 1 A
B 1 B
C 1 C
D 1 D
E 1 E"



for(j in 1:nsim){
    file <- paste("control_jc",j,sep="")
    treefile <- paste("genetrees_jc",j,sep="")
    b<-floor(runif(1)*619136+431171)
    a<-c(treefile,"0", b, paste(ngene,nspecies), c ,"0")
    write(a, file)
}
# gives control file for JC



for(j in 1:nsim){
    file <- paste("control_hky",j,sep="")
    treefile <- paste("genetrees_hky",j,sep="")
    b<-floor(runif(1)*619136+431171)
```

```
    a<-c(treefile,"0", b, paste(ngene,nspecies), c ,"0")
    write(a, file)
}
# gives control file for HKY


for(j in 1:nsim){
    file <- paste("control_gtr",j,sep="")
    treefile <- paste("genetrees_gtr",j,sep="")
    b<-floor(runif(1)*619136+431171)
    a<-c(treefile,"0", b, paste(ngene,nspecies), c ,"0")
    write(a, file)
}
# gives control file for GRT


for (j in 1:nsim){
    runfile<-paste("run_m",j,sep="")
    x<-c(paste("mpest control_jc",j,sep=""), paste("mpest control_hky",j,sep=""), paste(
    write(x,runfile)
}
# gives run_m files


write(paste("qsub -q rcc-30d run_m",1:nsim,sep=""),"submit_m")
# gives submit_m file
```

```
#submit jobs through submit_m
# give .tre files

nsim<-100
ngenetrees<-100

truegenetree<-1:(nsim*ngenetrees)
for(i in 1:(nsim*ngenetrees))
{
    truegenetree[i]<-read.tree.string(paste("truegenetree",i,sep=""),format="phylip")$tr
}
genetree_jc<-matrix("",ngenetrees,nsim)
genetree_hky<-matrix("",ngenetrees,nsim)
genetree_gtr<-matrix("",ngenetrees,nsim)
for(i in 1:nsim)
{
    genetree_jc[,i]<-scan(paste("genetrees_jc",i,sep=""),what="character",sep="\n")
    genetree_hky[,i]<-scan(paste("genetrees_hky",i,sep=""),what="character",sep="\n")
    genetree_gtr[,i]<-scan(paste("genetrees_gtr",i,sep=""),what="character",sep="\n")
}


jcdist <-1:length(genetree_jc)
hkydist <-1:length(genetree_hky)
gtrdist <- 1:length(genetree_gtr)
```

```
 for(k in 1:length(genetree_jc))
{
jcdist[k]<-dist.topo(read.tree(text=truegenetree[k]),read.tree(text=genetree_jc[k]))
hkydist[k]<-dist.topo(read.tree(text=truegenetree[k]),read.tree(text=genetree_hky[k]))
gtrdist[k]<-dist.topo(read.tree(text=truegenetree[k]),read.tree(text=genetree_gtr[k]))

}

 jcdata <- c(mean(jcdist),sd(jcdist))
 hkydata <- c(mean(hkydist),sd(hkydist))
 gtrdata <- c(mean(gtrdist), sd(gtrdist))
```

Bibliography

Darriba, D., Taboada, G., Doallo, R., and Posada, D. (2012). jmodeltest 2: more models, new heuristics and parallel computing. *Nature Methods*, 8:772.

Felsenstein, J. (1981). Evolutionary trees from dna sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, 17:368–376.

Guindon, S., Dufayard, J., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of phyml 3.0. *Systematic Biology*, pages 307–321.

Guindon, S. and Gascuel, O. (2003). A simple, fast and accurate method to estimate large phylogenies by maximum-likelihood. *Systematic Biology*, pages 696–704.

Jennings, W. B. and Edwards, S. V. (2005). Speciational history of australian grass finches (poephila) inferred from thirty gene trees. *Evolution*, 59:2033–2047.

Keast, A. (1958). Infraspecific variation in the australian finches. *Comparative Zoology*, 58:219–246.

Liu, L., Pearl, D., Brumfield, R., and Edwards, S. (2008). Estimating species trees using multiple-allele dna sequence data. *Evolution*, 62:2080–2091.

Posada, D. and Buckley, T. (2004). Model selection and model averaging in phylogenetics: Advantages of akaike information criterion and bayesian approaches over likelihood ratio tests. *Systematic Biology*, pages 793–808.

Rannala, B. and Yang, Z. (2003). Bayes estimation of species divergence times and acestral population sizes using dna sequences from multiple loci. *Genetics*, 164:1645–1656.

Smarda, P., Bures, P., Horova, L., Leitch, I., Mucina, L., Pacini, E., and Rotreklova, O. (2014). Ecological and evolutionary signficanced of genomic gc content diversity in monocots. *Proceedings of the National Academy of Science of the United States of America*, 111:E4096–E4102.

Yxuhehybyja (2010). Sequence alignment. `http://www.bioinformatics.org/wiki/Sequence_alignment`. Accessed: 2017-03-12.