Comparing models of molecular evolution for phylogenetic analysis of the scale worm family Polynoidae

Jason Moggridge 1159229

10/12/2020

Contents

This work is available at github.com/jmoggridge/6210_project5_phylogeny

The following software libraries were used in this work:

```
library(tidyverse)
library(bold) # data source api
library(Biostrings) # for sequence handling (*clashes w dplyr)
library(ape) # general phylogeny functions, eg. dist.dna
library(DECIPHER) # for alignment, etc.
library(phangorn) # ML phylogeny, model testing
library(dendextend) # dendrograms, tanglegrams
# library(phylogram) # loaded later for dendrogram
library(rcartocolor) # non-ugly discrete color pals
library(ggthemes) # <3 tufte theme for ggplot <3
library(patchwork) # easy plot assembly using +,/,()
```

1. Introduction

Phylogenetic analysis is common in biodiversity studies and environmental assessment, however it can be far from clear how to proceed. Not only are there are various phylogenetic marker genes (or combinations thereof) to assess, types of alignment (nucleic acid or protein, symbolic or structural) and methods for estimating phylogeny (distance, maximum-parsimony, maximum-likelihood). In maximum-likelihood, molecular evolution is modeled as a Markov process (*ie.* memoryless) with parameters for the base frequencies and rates of substitution. More complex models account for different substitution rates and base frequencies, *eg.* general time-reversible (GTR). Additional models describe gamma-distributed variance in rates across sites (+G) and invariant sites (+I).

Maximum likelihood-based (ML) phylogenetic clustering methods optimize model parameters to maximize the probability of the data given the model (Felsenstein 1981). As no model is a perfect representation of nature, we need to apply some measure of goodness-of-fit (log-likelihood, AIC, BIC) to evaluate competing models (Posada and Crandall 2001; Posada 2008). Perhaps this model selection step is unnecessary and we can generally use the complex model GTR+I+G, as we are not particularly interested in the distances, but rather the phylogenetic inferences that we make based on them (Abadi et al. 2019).

In this work, I sought to explore the impact of model choice in whether model testing is worthwhile. Specifically, I wanted to find out whether different distance models would lead to altered phylogenetic trees. For this, I performed model testing to evaluate AIC, BIC and likelihood criteria, pairwise statistical tests of log-likelihood ratios and tree distances, and a visual comparison of two maximum-likelihood phylogenies in a tanglegram. I chose to use cytochrome c oxidase subunit I sequences of Polynoidae as a model dataset. Polynoidae are a large and speciose family of scale worms that have colonized challenging habitats in the deep sea, including sea caves and hydrothermal vents species (Zhang et al. 2017). New species are consistently found in surveys and determining their phylogenetic relationships to known species is of interest (Hatch et al. 2020).

---

2. Polynoid DNA sequence dataset acquisition

```
# Download BOLD specimen + seq data for taxon 'Polynoidae'
polynoids.bold <- bold::bold_seqspec(taxon = 'Polynoidae')
polynoids.bold <- polynoids.bold %>%
  select(processid, contains('name'), contains('marker'), nucleotides) %>%
  select(-c(trace_names, phylum_name, class_name, subspecies_name,
            marker_codes))
write_rds(polynoids.bold, 'polynoid.raw.rds', compress = 'gz')
```

Polynoidae sequence data were downloaded from the public Barcode of Life Database (BOLD) api on 2020-12-12, using their `bold` package. There were 1850 specimen records for this taxon in the database in total. The data of interest (identifiers, sequences, and taxonomic names) were saved as `polynoid.raw.rds`.

3. Data quality control and filtering, sample selection, and multiple sequence alignment

Functions for filtering and EDA plot

```
## Selection of data for analysis:
trim_and_filter_seqs <- function(bold.df, markers, minlen,
                                 maxlen, Nthreshold){
  # keep seqs of marker gene; trim N and gaps;
  # filter by length and Ns proportion
  bold.df <- bold.df %>%
    filter(!is.na(nucleotides) & markercode %in% markers) %>%
    mutate(seq = str_remove_all(nucleotides, '\\s|-|^N+|N+$'),
           seqlen = nchar(seq),
           N.prop = str_count(seq, 'N')/seqlen) %>%
    filter(seqlen >= minlen & seqlen <= maxlen & N.prop <= Nthreshold)
  return(bold.df)
}
EDA_plot <- function(bold.df, label) {
  # plot sequence lengths, GC%, species per genus
  a <- ggplot(bold.df, aes(x = seqlen)) +
    labs(x = 'Sequence length', subtitle = label) +
    geom_histogram() + geom_rangeframe() + theme_tufte()
  b <- bold.df %>%
    mutate(`% GC` = str_count(seq, '[GC]')/seqlen*100) %>%
    ggplot(aes(x=`% GC`)) +
    geom_histogram() + geom_rangeframe() + theme_tufte()
  c <-bold.df %>%
    group_by(genus_name) %>%
    summarize(species = length(unique(species_name))) %>%
    ggplot(aes(x=species)) +
    xlab('Species per genus') +
    geom_histogram() + geom_rangeframe() +theme_tufte()
  return(a|b|c)
}
```

Selection of *cytochrome c oxidase subunit 1* sequences & QC

```
# read in data & tidy
polynoids.bold <- read_rds('polynoid.raw.rds') %>%
  mutate(across(where(is.character), ~na_if(.x, ''))) %>%
  mutate(across(where(is.character), as_factor)) %>%
  mutate(nucleotides = as.character(nucleotides))
# remove any specimens missing genus & species names
polynoids.bold <- polynoids.bold %>%
  filter(!is.na(genus_name) & !is.na(species_name))
# check out all COI sequences (not shown)
coi.eda.all <- polynoids.bold %>%
  trim_and_filter_seqs(markers = 'COI-5P', minlen = 0,
                       maxlen = 1000, Nthreshold = 1) %>%
  EDA_plot('all COI')
summary(str_count(polynoids.bold %>%
                    trim_and_filter_seqs(markers = 'COI-5P', minlen = 500,
                                         maxlen = 800, Nthreshold = 1) %>%
                    pull(seq),'N'))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00000 0.00000 0.00000 0.01906 0.00000 4.00000
```

```
# apply filtering steps
polynoids.df <- polynoids.bold %>%
  filter(!is.na(genus_name) & !is.na(species_name)) %>%
  trim_and_filter_seqs(markers = 'COI-5P', minlen = 630,
                       maxlen = 680, Nthreshold = 0.05)
coi.eda.filtered <- EDA_plot(polynoids.df, label = 'filtered COI')
```

```
# tidy_bold_data() %>%
# Sample 1 sequence per species for analysis, create labels
set.seed(1)
polynoids.sample <- polynoids.df %>%
  group_by(species_name) %>%
  sample_n(1) %>%
  data.frame() %>%
  mutate(seqlabel = paste(genus_name, row_number())) %>%
  mutate(across(where(is.factor), fct_drop)) # drop empty factor levels
# Exploratory analysis plot
coi.sample.eda <- EDA_plot(polynoids.sample,
         paste('Sampled data: n =', nrow(polynoids.sample)))
```

Multiple sequence alignment

```
do_alignment <- function(sample.df, names_vector, browse=FALSE, verbose = FALSE){
  # Format and label seqs, reorient reverse complements, align
  seqs <- Biostrings::DNAStringSet(sample.df$seq)
  names(seqs) <- names_vector
  seqs <-  DECIPHER::OrientNucleotides(seqs, verbose = verbose)
  seqs.aligned <- DECIPHER::AlignSeqs(seqs, verbose = verbose)
  if (browse==TRUE){
    BrowseSeqs(seqs.aligned)
    }
  return(seqs.aligned)
}
# Alignment
polynoids.align <- do_alignment(polynoids.sample, polynoids.sample$seqlabel)
polynoids.bin <- as.DNAbin(polynoids.align)
rm(do_alignment)
```

Only Polynoidae specimens having complete taxonomic information and *cytochrome c oxidase subunit I* (CO1) DNA sequences of appropriate length (630-680 bp) with no more than 2% ambiguous bases were retained. This filtered dataset has 959 specimens from 38 genera, with 95 species in total. A single representative from each species was chosen at random for alignment and phylogenetic analysis. Sequence properties (length, GC-content) and species distribution among genera were examined during data selection (*eg.* fig. 1). A multiple sequence alignment was performed and checked after sequence orientation.

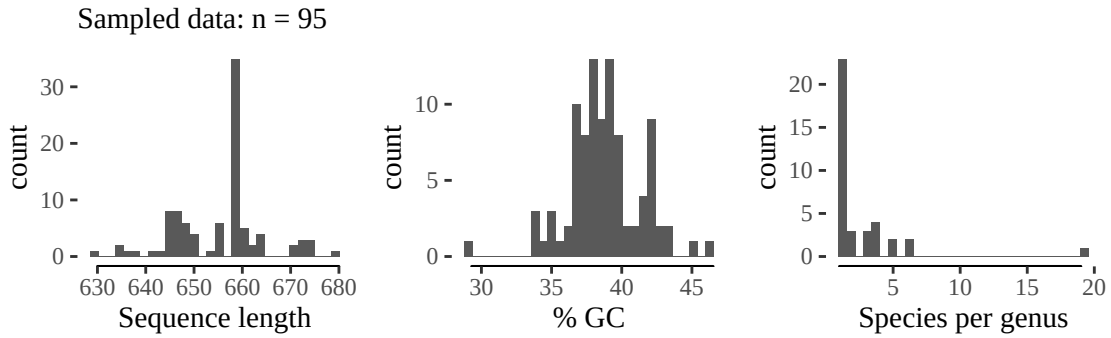

Figure 1. Exploratory analysis of sequence length (l), GC composition (c), and species distribution among genera (r) in the polynoid *cytochrome c oxidase subunit 1* sequences dataset used in this analysis (with one sequence per species).

---

4. Tools for model tests and phylogenetic analysis

This work relied heavily on the `phangorn` package for maximum-likelihood phylogenetic inference and model tests (Schliep 2011). Many R packages are available for ML-based phylogenetics (*eg.* RAxML), however `phangorn` has a large selection of models of molecular evolution (I used JC, F81, K80, HKY, SYM, GTR, though there are more) and statistical tests that were well-suited to this analysis; furthermore, any of the substitution models can be paired with +G (gamma-distributed substitution rate variance across sites) or +I (proportion of invariant sites) models, which was not the case with some other packages (eg. `DECIPHER` has no +I and needs to get with the times). Sequence alignment was done with the package `DECIPHER` (Wright 2016), as well as ML phylogeny (supplemental). The `ape` package (Paradis 2012) was used to compute some sequence distances; `dendextend` was used for dendrograms and tanglegrams. For the most part, all these analyses are well-demonstrated in the package vignettes and documentation (see references); this work extends those examples with pairwise statistical tests and tree distance tests for all models, with heatmaps for visualization (`ggplot2`).

---

5. Model testing and phylogenetic analysis

Molecular evolution model testing with `phangorn::modelTest`

```
# need phyDat format for test
polynoids.phy <- as.phyDat(polynoids.bin)
# calculate dist, make basic tree, then test subs. models
polynoids.dist <- phangorn::dist.ml(polynoids.phy, model='JC69')
polynoids.nj <- NJ(polynoids.dist)
# test default models selection; trace 0 makes it silent
model_test <- phangorn::modelTest(
  polynoids.phy, polynoids.nj, control = pml.control(trace=0)
  )
rm(polynoids.phy, polynoids.dist, polynoids.nj)
```

```
get_all_models <- function(model_test_obj){
  # get models from phangorn modelTest output using names list
  extract_model <- function(name, env){
    return(eval(get(name, env), env=env)) # retrieves model from model test output
  }
  tests.env <- attr(model_test_obj, "env")
  model.names <- ls(env=tests.env)[2:25]
  models <- lapply(model.names, extract_model, tests.env)
  names(models) <- model.names
  return(models)
}
get_models_pw_df <- function(models){
  # get_models_pw_df makes long-form df of pairwise matrix of models#
  m.names <- names(models)
  m.len <- length(m.names)
  # create long table of models, pairwise
  models.df <- tibble(
    name1 = rep(m.names, each = m.len),
    name2 = rep(m.names, times = m.len),
    model1 = rep(models, each = m.len),
    model2 = rep(models, times = m.len),
  ) %>%
    # no self-comparison; (make diagonals blanks)
    filter(!name1==name2) %>%
    rowwise() %>%
    mutate(tree1 = list(pluck(model1, 'tree')),
           tree2 = list(pluck(model2, 'tree'))) %>%
    # compute tree dists: Robinson-Foulds, weightedRF, branch-score diff, path diff
    mutate(
      RF.dist = phangorn::RF.dist(tree1, tree2, check.labels = TRUE),
      wRF.dist = phangorn::wRF.dist(tree1, tree2, check.labels = TRUE),
      KF.dist = phangorn::KF.dist(tree1, tree2, check.labels = TRUE),
      path.dist = phangorn::path.dist(tree1, tree2, check.labels = TRUE)
    ) %>%
    ungroup()
  return(models.df)
}
# get models, make a df, join model scores
models <- get_all_models(model_test)
models.df <- enframe(models) %>%
  left_join(., model_test, by = c('name' = 'Model'))
# just want names and AICc scores
AICc_scores <- model_test %>% select(Model, AICc)

# get all the models and perform treedist tests; calc deltaAIC
models.pairwise.df <- get_models_pw_df(models) %>%
  left_join(AICc_scores, by = c('name1' = 'Model')) %>%
  dplyr::rename(AICc.1 = AICc) %>%
  left_join(AICc_scores, by = c('name2' = 'Model')) %>%
  dplyr::rename(AICc.2 = AICc) %>%
  mutate(deltaAICc = AICc.2 - AICc.1)
```

Log-likelihood ratio test and tree distance metrics for HKY+G+I versus GTR+G+I

```
# grab best model from test, HKY+G+I
HKY_IG.phyl <- models.pairwise.df %>%
  filter(name1 == "HKY+G+I") %>%
  pull(model1) %>% pluck(1)
# grab most parameter-rich model, GTR+G+I
GTR_IG.phyl <- models.pairwise.df %>%
  filter(name1 == "GTR+G+I") %>%
  pull(model1) %>% pluck(1)
# do log-likelihood ratio test -> not significant!
anova(HKY_IG.phyl, GTR_IG.phyl)
```

```
## Likelihood Ratio Test Table
##   Log lik.  Df Df change Diff log lik. Pr(>|Chi|)
## 1   -20741 193
## 2   -20810 197         4       -137.99          1
```

```
# check topological dist -> the only difference is branch lengths!
treedist(HKY_IG.phyl$tree, GTR_IG.phyl$tree)
```

```
##      symmetric.difference   branch.score.difference           path.difference
##                 0.0000000                 0.6786923                 0.0000000
## quadratic.path.difference
##                17.9316038
```

---

6. Figures

```
# plotted in fig2: logLik, AICc, and BIC; model deg.freedom
model_test_plot <- model_test %>%
  select(Model, AICc, BIC, logLik, df)  %>%
  pivot_longer(AICc:df) %>%
  mutate(name = as_factor(name)) %>%
  ggplot(aes(x=Model, y=value, colour = name)) +
  geom_point() +
  labs(x = '', y = '') +
  facet_grid(name~., scales = 'free_y') +
  scale_color_carto_d() + theme_light() +
  theme(legend.position = 'none',
        axis.text.x = element_text(angle = 90, hjust = 1, vjust = 0.5),
        axis.text.y = element_text(size = 7)
  )
model_test_plot
```

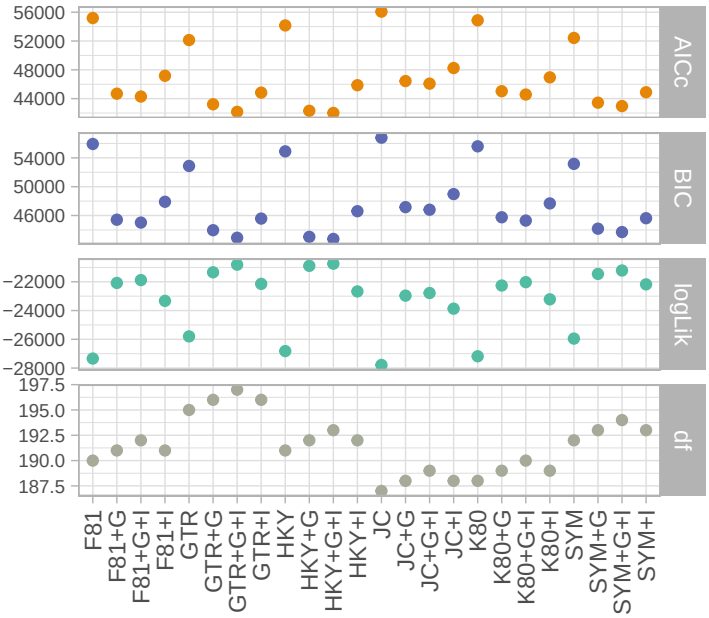

Figure 2. Testing of all DNA substitution models (JC, F81, K80, HKY, TrN, TPM, K81, TIM, TVM, SYM, GTR) with and without added site-specificity (G, I) for the polynoid CO1 sequences, using `phangorn::modelTest`. The AIC, BIC and log likelihood scores (y-axis) are shown for each model (x-axis), where smaller AICc and BIC scores and larger log likelihood scores indicate better fit. The degrees of freedom for each model are included (df).

```
# Create heatmap for log-likelihood difference
a <- ggplot(models.pairwise.df, aes(y=name2, x=name1, fill=deltaAICc)) +
  geom_tile(na.rm = TRUE) +
  scale_fill_viridis_c(option='D',direction = -1) +
  theme_light() +
  labs(x='', y='',
       fill='AICc(y) - AICc(x)') +
  theme(axis.text.x = element_text(angle = 90, hjust = 1, vjust = 0.5),
        legend.position = 'left')
# Create heatmap for weighted Robinson-Foulds distance
b <- models.pairwise.df %>%
  # keep only 'upper triangle' if df was sq matrix
  mutate(index1 = match(name1, names(models)),
         index2 = match(name2, names(models))) %>%
  filter(!index2 <= index1) %>%
  select(-index1, -index2) %>%
  ggplot(aes(y=name2, x=name1, fill=wRF.dist)) +
  geom_tile(na.rm = TRUE) +
  scale_fill_viridis_c(option='C') +
  theme_light() +
  labs(x='', y='',
       fill='wieghted RF distance') +
  theme(axis.text.x = element_text(angle = 90, hjust = 1, vjust = 0.5),
        legend.position = 'top')
(a+b)
```

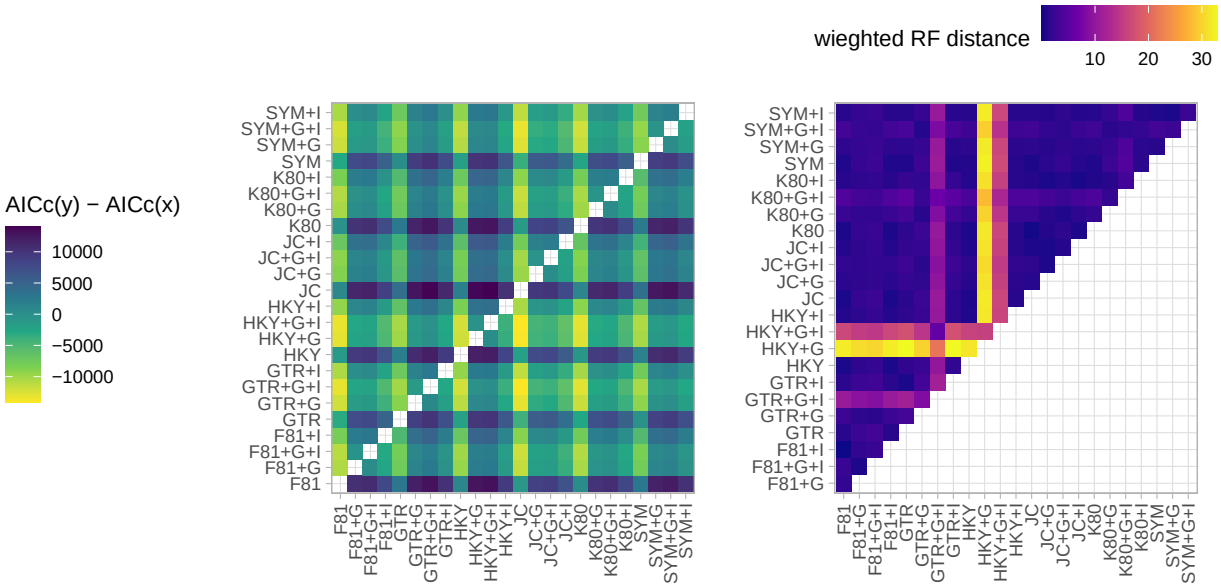

Figure 3. Heatamaps for the difference in AIC with small-samples correction (AICc) between each pair of models (left), and weighted Robinson-Foulds topological distance from pairwise comparisons all models tested using **phangorn** (right). Note that smaller AICc scores indicate better fit, so rows that have smaller delta(AICc) values represent better models (and the opposite is true for columns), where deltaAICc = AICc(model.y) - AICc(model.x).

```
library(phylogram)
# need 'phylogram' library to convert class phylo to dendrogram
HKY_dendrogram <- as.dendrogram(HKY_IG.phyl$tree) %>%
  set('labels_cex', 0.4) %>%
  color_branches(k=38)
ggplot(HKY_dendrogram, horiz = TRUE)
```

Figure 4. Dendrogram of the maximum-likelihood phylogeny with the HKY+G+I model.

7. Supplemental

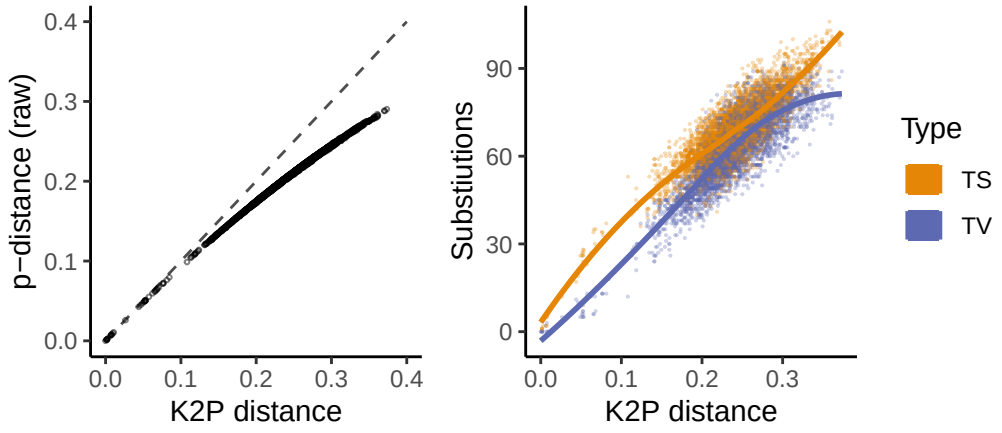

Figure 5. Saturation plots: p-distance underestimates real evolutionary distance as distance increases, while the Kimura 2-parameter is more robust to the saturation of transitions (left). The proportion of transitions to is generally greater that transversions amongst the family Polyploidae, where phylogenetic distances are relatively small and transitions are not entirely saturated (right); if we were to compare all polychaete species, for example, the lines would cross and transversions would be more common as transitions become saturated.

DECIPHER has no implementation of the GTR+I+G model, so I wondered how it's chosen model would compare. TN93+G(4) was chosen and the scores were very similar to the HKY+G+I model generated in the `phangorn` model testing. I also wanted to see how different the phylogenies would be compared a distance-based method with the same substitution model.

```
## Maximum-likelihood hierarchical clustering
clusters.ML <- DECIPHER::IdClusters(
  myXStringSet = polynoids.align,
  myDistMatrix = DistanceMatrix(polynoids.align, type='dist'),
  method = 'ML',  cutoff = 0.2, showPlot = FALSE, type = "both", verbose = FALSE)


## ================================================================================
##
## Time difference of 0.05 secs

clusters.Dist <- DECIPHER::IdClusters(
  myXStringSet = polynoids.align,
  myDistMatrix = DistanceMatrix(polynoids.align, type='dist'),
  model = 'TN93', method = 'NJ', cutoff = 0.2, showPlot = FALSE, type = "both", verbose = FALSE)


## ================================================================================
##
## Time difference of 0.01 secs
```

```
# library(phylogram)
dend.list <- dendlist(
    clusters.Dist[[2]] %>% set('labels_cex', 0.65) %>%
      set("branches_lwd", 1),
    clusters.ML[[2]] %>% set('labels_cex', 0.65) %>%
      set("branches_lwd", 1)
    ) %>%
  dendextend::untangle() %>%
# Make tanglegram
  tanglegram(
    main_right = 'ML, TN93+G model', main_left = 'dist-based, TN93 model',
    match_order_by_labels = TRUE, just_one = TRUE,
    columns_width = c(6, 1, 6), left_dendo_mar = c(2, 2, 2, 6),
    right_dendo_mar = c(2, 6, 2, 2)
  )
```

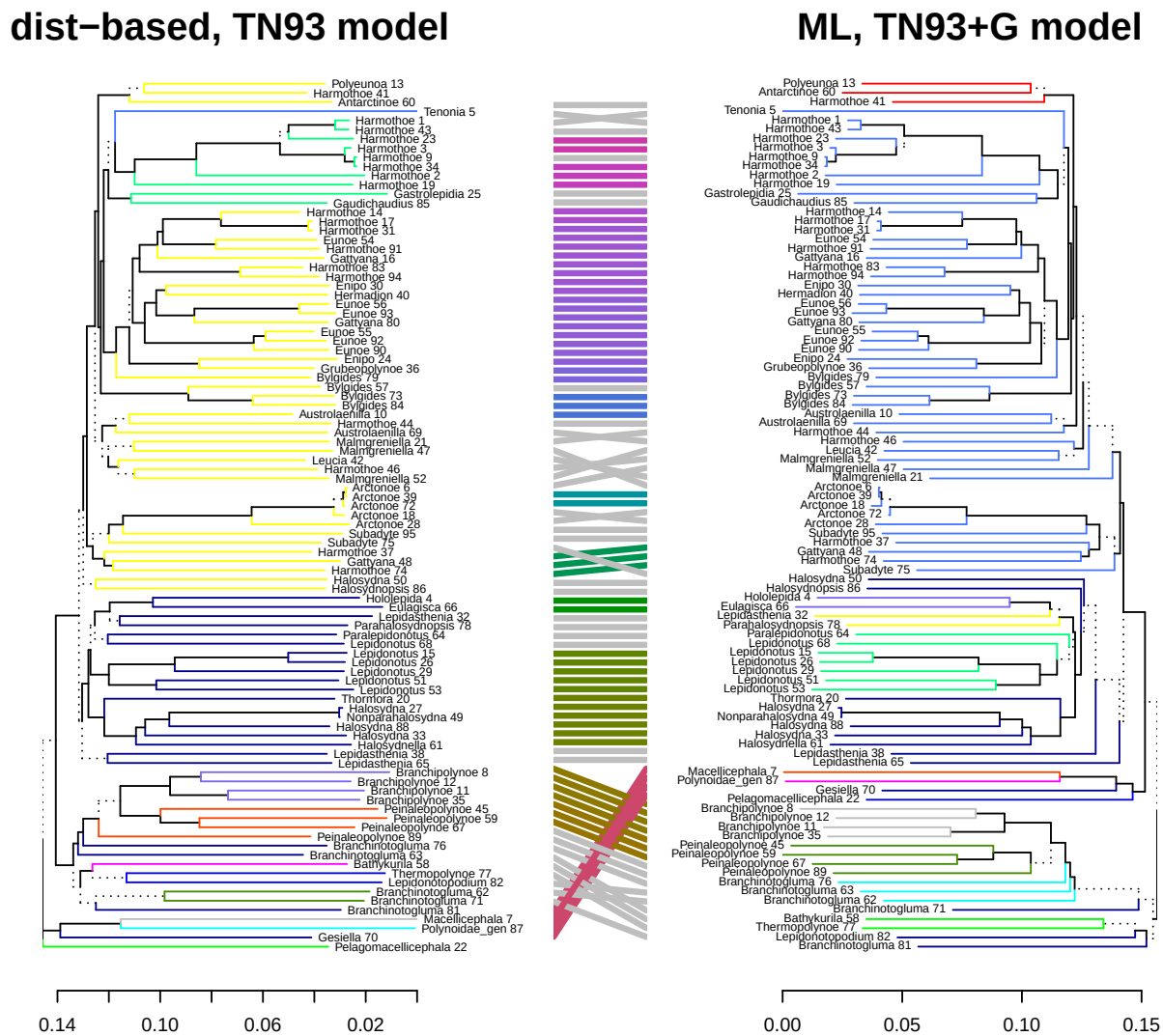

Figure 6. A tanglegram showing differences between a distance-based phylogeny using the Jukes-Cantor model and a ML phylogeny using the TN93+G model, selected based on the Bayesian information criteria (BIC).

8. Results and Discussion

There is disagreement in the literature as to whether model selection significantly impacts maximum likelihood (ML) phylogenetic inferences (Abadi et al. 2019; Gerth 2019). In this project, I wanted to see if model selection was in fact a crucial step in an ML phylogenetic analysis of the family Polynoidae, using CO1 DNA sequences. In model testing HKY+I+G was the best performing model across all key criteria (figs. 2, 3-left); however, differences in goodness-of-fit between this model and the most parameter-rich, GTR+G+I, were not significant in a likelihood ratio test. Interestingly, the incorporation of rate variation among sites (+G), and static sites (+I) had much more impact that differences between various substitution models (fig.3-left). Further, the model choice did not affect the partitioning of polynoid species whatsoever (according to the Robinson-Folds RF distance), though when branch lengths were accounted for (in weighted RF), the HKY+G, HKY+G+I and GTR+G+I models were slightly different to all other models tested (fig. 3-right). Thus, my results indicated that most parameter-rich models will produce very similar phylogenies and that model testing was unnecessary. This was not unexpected, as Abadi *et al.* (2019) have suggested that ML model testing generally does not greatly impact inferences, in contravention with *eg.* Posada *et al.* (2008).

Avoiding model testing is most beneficial when datasets are large and optimization of many models would be time-consuming and computationally intensive. Additionally, simpler models would generally only be preferable to the more realistic GTR+I+G when the data are limited, leading to greater expected error in parameter estimation. While my results agree that model testing generally seems unnecessary, this project only used a single, small dataset (n=95 species), of a single region (CO1), and only for DNA sequences (fig. 4); this dataset was drawn from a single speciose family where distances are relatively small (tree height ~ 0.2, depending on model) and was chosen mostly for convenience and tractability. Furthermore, some available models were not tested at all for brevity (TIM, TVM) and only a single algorithm, `phangorn::modelTest` (Schliep 2011), was used for testing, though many other similar tools could have been employed for comparison. Perhaps inclusion of all models or other tools could have uncovered a model with significantly better fit or producing different topology. To determine whether model choice impacts phylogenetic inferences in general would require analysis of multiple, varied datasets and this is generally done through simulated data with known topology to evaluate concordance (Abadi et al. 2019). As such, my results should only be interpreted in the limited context of the Polyploidae CO1 DNA sequences analyzed and the selection of models that were tested.

Model testing is but one step in ML phylogenetic analysis, albeit a particularly time-consuming one. Many other choices could have a far greater impact on inferences, for example: the selection of phylogenetic marker(s); use of DNA sequences versus their more conserved translations; inclusion of structural information, in the case of rRNAs; alignment parameters, etc. A more complete version of my analysis would have compared the utility of model selection in these various scenarios; this could provide further support to the recent proposition that model selection is unimportant for ML phylogeny estimation (Abadi et al. 2019), or it's rebuttal (Gerth 2019). This could be particularly helpful in the context of analysis of high-throughput sequencing data, where model selection may represent a major burden upon computational resources.

By conducting this analysis, I gained insight into the details that should be considered when applying ML phylogeny methods to real biological data. While the scope of this project was fairly limited, I was exposed to many different methods aside from those applied here. Phylogenetic analysis in the R environment was a bit daunting when starting out; though over time, I developed a grasp of the statistical underpinnings of maximum likelihood methods, how a work-flow could be structured, which tools are commonly applied, how to evaluate different models, and which additional packages are available for specific tasks such as visualization. This will undoubtedly be of benefit when performing phylogenetic analyses in the future.

---

9. Acknowledgements

I would like to thank Dr. Sarah Adamowicz and Jacqueline May for the advice and support they provided in developing ideas for this analysis and suggestions about which packages and tools might be useful.

10. References

In preparing this work, I found the following tutorials, vignettes, and documentation most helpful:

- Estimating phylogenetic trees with phangorn, K.P. Schliep
- Phagorn specials vignette, K.P. Schliep
- Getting the models from the output of `phangorn::modelTest`
- The art of multiple sequence alignment in R by E.S. Wright
- Introduction to `dendextend` by T. Galili
- Genetic data analysis using R: introduction to phylogenetics by T. Jombart

---

Books and Journal Articles

- *NB: Any articles not available through the University of Guelph subscriptions were shamelessly pirated through sci-hub.tw*

Abadi, Shiran, Dana Azouri, Tal Pupko, and Itay Mayrose. 2019. "Model selection may not be a mandatory step for phylogeny reconstruction." *Nature Communications* 10 (1). https://doi.org/10.1038/s41467-019-08822-w.

Felsenstein, Joseph. 1981. "Evolutionary trees from DNA sequences: A maximum likelihood approach." *Journal of Molecular Evolution* 17 (6): 368–76. https://doi.org/10.1007/BF01734359.

Gerth, Michael. 2019. "Neglecting model selection alters phylogenetic inference." *bioRxiv*, 1–7. https://doi.org/10.1101/849018.

Hatch, Avery S., Haebin Liew, Stéphane Hourdez, and Greg W. Rouse. 2020. "Hungry scale worms phylogenetics of peinaleopolynoe (Polynoidae, annelida), with four new species." *ZooKeys* 2020 (932): 27–74. https://doi.org/10.3897/zookeys.932.48532.

Paradis, Emmanuel. 2012. *Analysis of Phylogenetics and Evolution with R*. 2nd ed. Springer. https://doi.org/10.1007/978-1-4614-1743-9_7.

Posada, David. 2008. "jModelTest: Phylogenetic model averaging." *Molecular Biology and Evolution* 25 (7): 1253–6. https://doi.org/10.1093/molbev/msn083.

Posada, David, and Keith A. Crandall. 2001. "Selecting the Best-Fit Model of Nucleotide Substitution." *Systematic Biology* 50 (4): 580–601. https://doi.org/10.1080/106351501750435121.

Schliep, Klaus Peter. 2011. "phangorn: Phylogenetic analysis in R." *Bioinformatics* 27 (4): 592–93. https://doi.org/10.1093/bioinformatics/btq706.

Wright, Erik S. 2016. "Using DECIPHER v2.0 to analyze big biological sequence data in R." *R Journal* 8 (1): 352–59. https://doi.org/10.32614/rj-2016-025.

Zhang, Yanjie, Jin Sun, Chong Chen, Hiromi K. Watanabe, Dong Feng, Yu Zhang, Jill M. Y. Chiu, Pei Yuan Qian, and Jian Wen Qiu. 2017. "Adaptation and evolution of deep-sea scale worms (Annelida: Polynoidae): insights from transcriptome comparison with a shallow-water species." *Scientific Reports* 7 (June 2016): 1–14. https://doi.org/10.1038/srep46205.