Clever caw-mparisons: examining different nucleotide substitution models in a Bayesian approach to corvid phylogeny

By Meilin Yen

Background

The Bayesian approach to phylogeny focuses on the joint posterior probabilities of the phylogenetic trees conditional on the alignment of the observed DNA sequences. The joint posterior probabilities can be found using Bayes' theorem:

$$f(t|X) = \frac{f(X|t)f(t)}{f(X)} \qquad (1)$$

where *X* represents the data and *t* represents a specific tree. Markov Chain Monte Carlo (MCMC) methods are used to approximate $f(X|t)$ and $f(t|X)$. $f(t|X)$ can be found through traditional MCMC methods, but estimating $f(X|t)$ requires the usage of a power posterior. Two common methods are path sampling (Gelman & Meng, 1998) and stepping-stone sampling (Xie et al., 2011).

The probability $f(X|t)$ depends on the nucleotide substitution model. The model determines *Q*, the instantaneous rate matrix used in the MCMC calculations. I will be examining the following models: the Jukes-Cantor model (JC69) (Jukes et al., 1969), the Felsenstein 1981 model (F81) (Felsenstein, 1981), the Hasegawa-Kishino-Yano 1985 model (HKY85) (Hasegawa et al., 1985), the General Time-Reversible model (GTR) (Tavaré, 1986), and the General Time-Reversible model with a Discrete Gamma (GTR + Γ) (Yang, 1994).

The family *Corvidae* is a charismatic clade of birds. Yet much is still unknown about their evolution, as our understanding of avian phylogeny constantly changes. Different single nucleotide substitution models could better explain different phylogenies. By comparing various nucleotide substitution models, I could determine which one best fits the *Corvidae* and learn more about their evolutionary history.

Past research determined that the GTR + Γ + Invariable sites model was best suited for maximum likelihood analyses, but the authors failed to mention what model was used in the Bayesian analysis (Ericson et al., 2005). Specific model comparison within the Bayesian approach for different genes had also been performed on New World Jays (Bonaccorso et al., 2007), a clade within *Corvidae*. Model selection analysis has not yet been done for the entire family.

Research Hypothesis

The null hypothesis is that there are no significant differences in model performance. The alternative hypothesis is that at least one of the models is significantly better at describing the data.

The Bayes factor quantifies differences in model performance by comparing conditional probabilities of observing the data given the model. The equation is

$$BF(M_0, M_1) = \frac{\Pr(X|M_0)}{\Pr(X|M_1)} = \frac{\Pr(M_0|X)}{\Pr(M_1|X)} \div \frac{\Pr(M_0)}{\Pr(M_1)} \qquad (2)$$

where *X* represents the data and M_0 and M_1 represent the two models being compared. If the Bayes factor is greater than 100, then M_0 is significantly better than M_1 (Jeffreys, 1961). Otherwise, the difference in performance between the two models is not large enough to be significant. Bayes factors less than 1 signify support in favor of M_1.

I am supplementing my Bayesian analysis with maximum likelihood analysis. The likelihood ratio test statistic (Equation 3) is used to compare two models' performances.

$$-2ln(\Lambda),\ \Lambda = \frac{max(L_0)}{max(L_1)} \qquad (3)$$

max(L$_0$*)* is the maximum likelihood for Model 0 and *max(L*$_1$*)* is the maximum likelihood for Model 1. To determine significance, compare $-2ln(\Lambda)$ to critical values from a Chi-squared distribution with degrees of freedom equal to the difference in degrees of freedom between Model 0 and Model 1. If $-2ln(\Lambda)$ is greater than some critical value, then Model 1 does significantly better than Model 0. Otherwise, there is no significant difference in performance.

I predict that the null hypothesis will be rejected and that the GTR + Γ model will do significantly better than all the others. This prediction means that the Bayes Factors where M_0 = GTR + Γ and $M_1 \neq$ GTR + Γ will all be greater than 100. The GTR + Γ model is the most complex model used in this analysis, which means that it can account for the highest amount of complexity in the data and will then be the "best" model.

Methods

Model and degrees of freedom	Parameters	Distributions
JC69 (0 degrees of freedom)	None	None
F81 (3 degrees of freedom)	Stationary frequencies (π)	Flat Dirichlet distribution with α = 1 and 4 elements in the Dirichlet random variable vector
HKY85 (4 degrees of freedom)	Stationary frequencies (π) Transition-Transversion rate (κ)	Flat Dirichlet distribution with α = 1 and 4 elements in the Dirichlet random variable vector Lognormal Distribution with mean 0 and standard deviation 1
GTR (8 degrees of freedom)	Stationary frequencies (π) Exchangeability rates (r)	Flat Dirichlet distribution with α = 1 and 4 elements in the Dirichlet random variable vector Flat Dirichlet distribution with α = 1 and 6 elements in the Dirichlet random variable vector
GTR + Γ (9 degrees of freedom)	Stationary frequencies (π) Exchangeability rates (r) Discrete gamma distribution shape parameter (α)	Flat Dirichlet distribution with α = 1 and 4 elements in the Dirichlet random variable vector Flat Dirichlet distribution with α = 1 and 6 elements in the Dirichlet random variable vector Uniform distribution from 0 to 10

Table 1: Comparisons of the models' degrees of freedom, the parameters in the models, and the corresponding distributions

The data consisted of cytochrome b (*CYTB*) gene sequences (1143 bp) from the mitochondrial genome for 29 species in the family *Corvidae* across 11 different genera. One outgroup, *Lanius ludovicianus* (family: *Laniidae*), was chosen due to the *Laniidae* family's close evolutionary relationship to the *Corvidae*. Recent research supported the idea that the *Eurocephalus* genus was a sister clade to the *Corvidae* (McCullough et al., 2023), but no data for this genus was available. A past study placed the *Laniidae* and the *Platylophidae* families as a sister clade to *Corvidae* (Kuhl et al., 2021), but no genomic data was available for any species in the *Platylophidae* family.

All genomic data was obtained from the National Center for Biotechnology Information (NCBI) online database. Four species included in the NCBI query were omitted from analysis because those species are not widely considered to be in *Corvidae*. The remaining species' sequences were aligned in one Nexus file using MEGA11 (Tamura et al., 2021).

MCMC and power posterior analyses were both performed in RevBayes (Höhna et al., 2016). One RevBayes script per type of analysis was created for each model. The MCMC analysis was used to find the maximum *a posteriori* non-clock phylogenetic trees. The power posterior analysis was used to find the average log likelihoods and included the path sampling and stepping-stone sampling methods. Each MCMC script's Markov chain ran for 100,000 iterations with a tuning interval of 100 iterations. Each power posterior script ran for 1,000 iterations with a burn-in period of 10,000 iterations and a tuning interval of 1000 iterations.

Scaling moves were used for branch lengths and NNI (Nearest Neighbor Interchange) moves (Felsenstein, 2004) were used for tree topologies. Table 1 contains the distributions used to generate parameter values for the instantaneous rate matrices. The moves for the parameters in the models corresponded to the parameters' distributions. Log maximum likelihoods for further analysis were obtained with assistance from Professor John Huelsenbeck using a heuristic search in PAUP* (Swofford, D. L., 2003). No invariable sites analysis was done.

The log Bayes factors were calculated using a Python script and the likelihood ratio test values were calculated by hand. By modifying Equation 2, I obtain the equation for the log Bayes factors.

$$ln(BF(M_0, M_1)) = ln(\frac{\Pr(X|M_0)}{\Pr(X|M_1)}) = ln(\Pr(X|M_0)) - ln(\Pr(X|M_1)) \qquad (4)$$

Results

Model	Average log marginal likelihood (path sampling)	Average log marginal likelihood (stepping-stone sampling)	Log maximum likelihood
JC69	-11061.46	-11062.07	-10874.650
F81	-10865.66	-10867.75	-10663.611
HKY85	-10235.17	-10239.37	-10009.600
GTR	-10113.58	-10115.56	-9870.523

GTR + Γ	-9125.033	-9125.816	-8889.061

Table 2: Average log marginal likelihoods (found using RevBayes) and log maximum likelihoods (found using PAUP*) for each model

	JC69	F81	HKY85	GTR	GTR + Γ
JC69	0	-195.8	-826.29	-947.88	-1936.427
F81	195.8	0	-630.49	-752.08	-1740.627
HKY85	826.29	630.49	0	-121.59	-1110.137
GTR	947.88	752.08	121.59	0	-988.547
GTR + Γ	1936.427	1740.627	1110.137	988.547	0

Table 3: Natural log Bayes factors for each combination of models, with each row representing M_0 and each column representing M_1 and average log marginal likelihoods calculated using the path sampling method

	JC69	F81	HKY85	GTR	GTR + Γ
JC69	0	-194.32	-822.7	-946.51	-1936.254
F81	194.32	0	-628.38	-752.19	-1741.934
HKY85	822.7	628.38	0	-123.81	-1113.554
GTR	946.51	752.19	123.81	0	-989.744
GTR + Γ	1936.254	1741.934	1113.554	989.744	0

Table 3: Natural log Bayes factors for each combination of models, with each row representing M_0 and each column representing M_1 and average log marginal likelihoods calculated using the stepping-stone sampling method

Model	Difference in degrees of freedom compared to GTR + Γ	Likelihood ratio test statistic
JC69	9	3971.18
F81	6	3549.1
HKY85	5	2241.078
GTR	1	1962.924

Table 5: Likelihood ratio test values for GTR + Γ compared to each of the other models

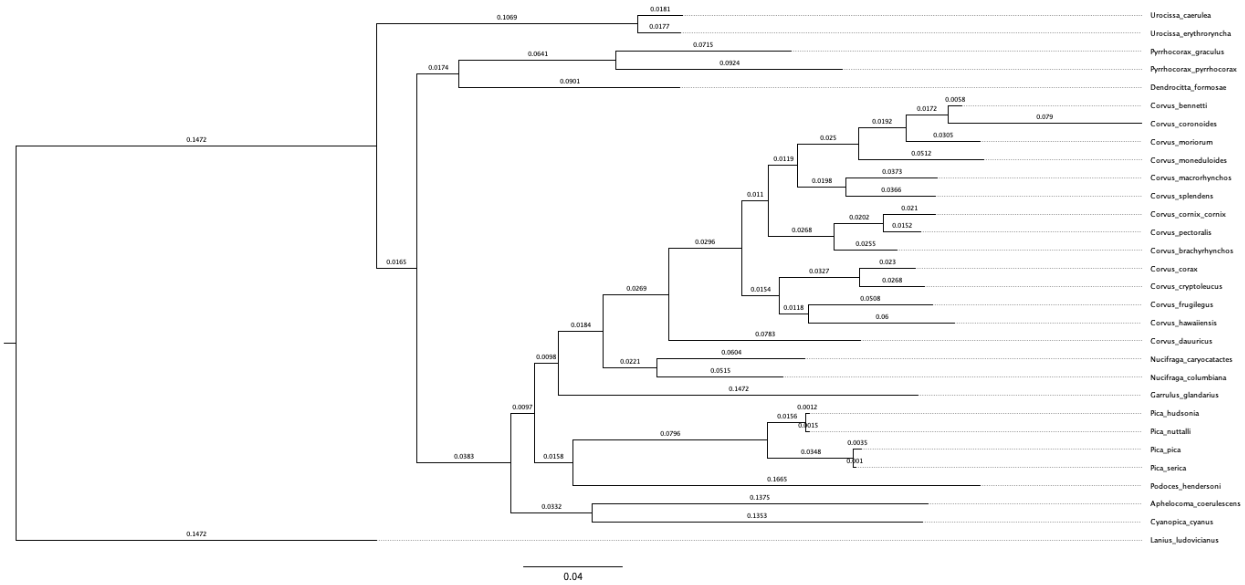


Figure 1: The maximum *a posteriori* non-clock tree generated underneath the GTR + Γ model, with branch length equal to expected number of substitutions per site, and the reference line being a branch of length 0.04 expected substitutions per site

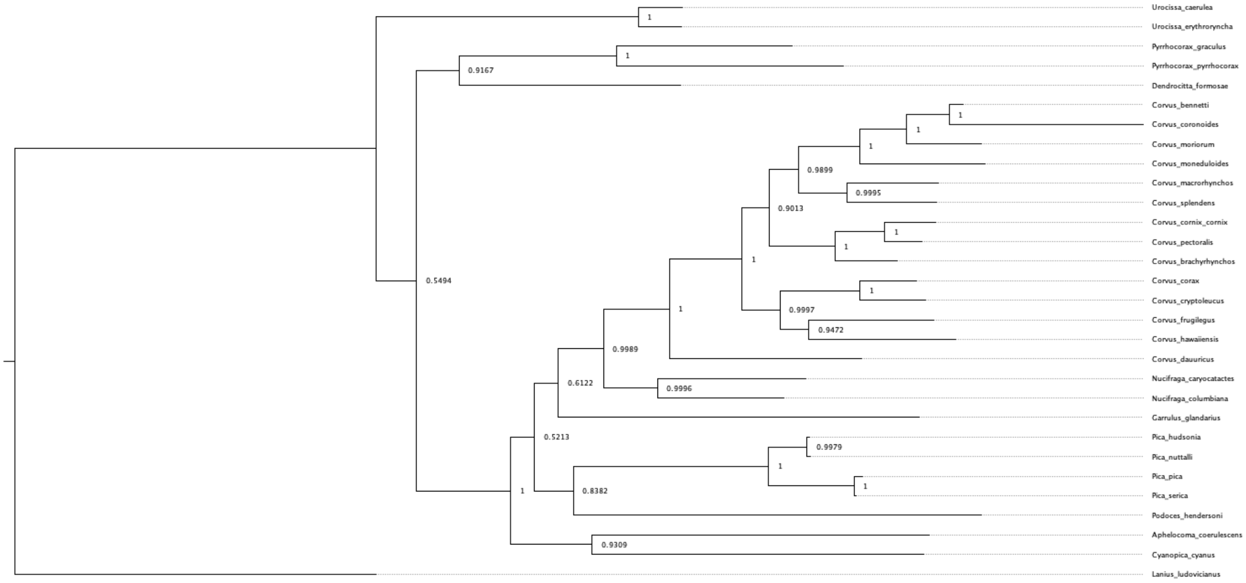
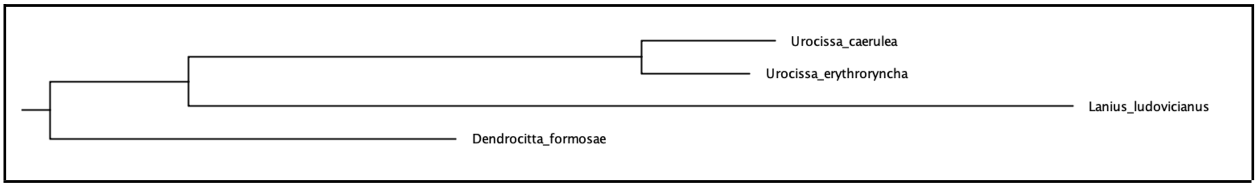


Figure 2: The maximum *a posteriori* non-clock tree generated underneath the GTR + Γ model, with node labels equal to the posterior probabilities for the clades being monophyletic under the model



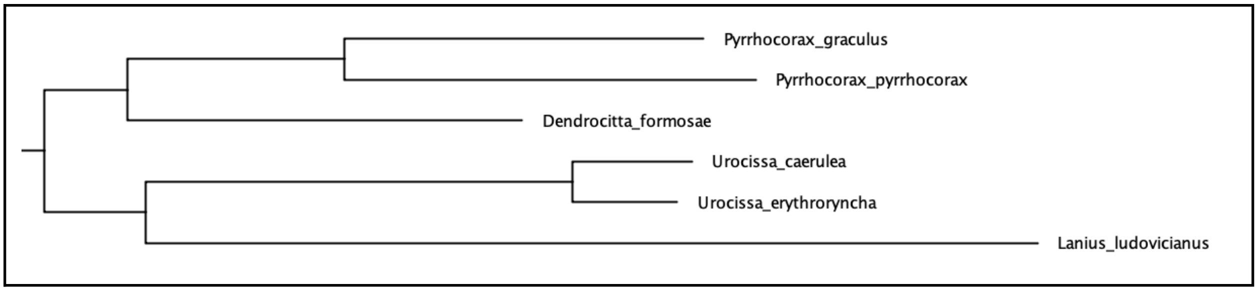


Figure 3: Smallest clades containing *D. formosae*, *L. ludovicianus*, *U. caerulea*, and *U. erythroryncha* in the JC69 maximum *a posteriori* tree (top) and the GTR maximum *a posteriori* tree (bottom)

The raw average log likelihood values obtained through the power posterior analysis are summarized in Table 2. Log likelihoods (base *e*) were calculated in order to avoid issues with numerical underflow. There were small differences in the average log likelihoods per model across the path sampling and stepping-stone sampling methods. For all the single nucleotide substitution models used, the log maximum likelihood (base *e*) was higher than the average log likelihoods for both sampling methods in the Bayesian analysis.

The absolute value of the log Bayes factors in Tables 3 and 4 increased as the difference in model complexities grew. Similarly, in Table 5, as the difference in degrees of freedom decreased, the likelihood ratio test statistic decreased as well.

With regards to the maximum *a posteriori* trees, the posterior clade probabilities and the branch lengths differed most often across models, although changes in topology appeared as well. For example, in the JC69 maximum *a posteriori* tree, the smallest clade containing *Dendrocitta formosae*, *Lanius ludovicianus*, *Urocissa caerulea*, and *Urocissa erythroryncha* (Figure 3), contained just those four species. Meanwhile, in the GTR maximum *a posteriori* tree, the smallest clade containing these four species (Figure 3) also included the *Pyrrhocorax* genus.

All maximum *a posteriori* trees were visualized using FigTree v1.4.4 (Rambaut, 2010).

Discussion

It is unclear what particular differences between the Bayesian and maximum likelihood approaches led to the difference in the average and maximum log likelihoods. However, in both approaches, the GTR + Γ model outperforms all other models. I can reject the null hypothesis.

If a log Bayes factor greater than 4.6 (ln(100) = 4.6) implies that there is decisive evidence in support of M_0 (Jeffreys, 1961), then the high log Bayes factors in the last rows of Tables 3 and 4 suggested that there was always decisive support in favor of M_0 = GTR + Γ. As a result, the maximum *a posteriori* tree generated under the GTR + Γ model and its associated expected number of substitutions per site (Figure 1) and clade posterior probabilities (Figure 2) also best reflect the phylogenetic relationships underlying the data.

The values from the log likelihood ratio tests, presented in Table 5, further support the results of the Bayesian analysis. The GTR + Γ model has the highest maximum log likelihood, but this model is also the most complex, so the log likelihood ratio tests were done to consider differences in complexity. Using $\alpha = 0.005$ for all degrees of freedom, all the likelihood ratio test statistics were greater than all of the corresponding critical values.

The GTR + Γ model, by virtue of being the most complex model in this analysis, can best fit highly complex avian genomic data. Unlike other models, the GTR + Γ model relies less on assumptions that are frequently false in nature, such as assuming that substitution rates are the same across sites. By relaxing this assumption, performance improved, as the addition of the discrete gamma to the GTR model yielded large log Bayes factors (988.547 for the path sampling method estimates and 989.744 for the stepping-stone sampling estimates).

I omitted invariable site analysis due to weird interactions with other parts of the RevBayes scripts that affected parameter estimation. In the future, I would extend my project to include other types of single nucleotide substitution models, more than 1 tree rearrangement move, and the aforementioned invariable sites analysis. I would also obtain data for multiple genes and from a wider variety of species to broaden my analysis and see how gene choice and species data availability influenced my phylogenies and my model selection process.

Data Availability

The code used in this project is available on GitHub. This archive contains the RevBayes and Python scripts, the necessary data to create and run these scripts (see readme for details: https://github.com/mmyen/clever-cawmparisons-ib134l), and enlarged versions of Figures 1 and 2. The PAUP* script was unavailable.

Acknowledgements

I thank Professors Rebecca Tarvin and John Huelsenbeck as well as David Tian for their invaluable guidance and insights on this project. I thank my peer reviewers for their feedback and my family for their support. Lastly, I thank the crows who hang out in my family's backyard for their shiny gifts and their endless entertainment.

References

Bonaccorso, E., & Townsend Peterson, A. (2007). A multilocus phylogeny of New World jay genera. *Molecular Phylogenetics and Evolution*, 42(2), 467–476.

Ericson, Per. G. P., Jansén, Al., Johansson, U. S., & Ekman, J. (2005). Inter-Generic Relationships of the Crows, Jays, Magpies and Allied Groups (Aves: Corvidae) Based on Nucleotide Sequence Data. *Journal of Avian Biology*, 36(3), 222-34.

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

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

Gelman, A., & Meng, Xl (1998). Simulating normalizing constants: from importance sampling to bridge sampling to path sampling. *Statistical Science*, 13(2), 163-185.

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

Höhna, S., Landis, M. J., Heath, T. A., Boussau, B., Lartillot, N., Moore, B. R., Huelsenbeck, J. P., & Ronquist, F. (2016). RevBayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. *Systematic Biology*, 65(4), 726-736.

Jeffreys, H. (1961). *Theory of Probability*. Oxford University Press.

Jukes, T. H., & Cantor, C. R. (1969). Evolution of Protein Molecules. In H. N. Munro (Ed.), *Mammalian Protein Metabolism* (pp. 21-132). Academic Press.

Kuhl, H., Frankl-Viches, C., Bakker, A., Mayr, G., Nikolaus, G., Boerno, S. T., Klages, S., Timmermann, B., & Gahr, M. (2021). An Unbiased Molecular Approach Using 3′-UTRs Resolves the Avian Family-Level Tree of Life. *Molecular Biology*, 38(1), 108-127.

McCullough, J. M., Hruska, J. P., Oliveros, C. H., Moyle, R. G., Andersen, M. J. (2023). Ultraconserved elements support the elevation of a new avian family, Eurocephalidae, the white-crowned shrikes. *Ornithology*, 140(3), 1-11.

Rambaut, A. (2010). *FigTree* (Version 1.4.4). Institute of Evolutionary Biology, University of Edinburgh.

Swofford, D. L. (2003). *PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods)* (Version 4). Sinauer Associates.

Tamura, K., Stecher, G., & Kumar, S. (2021). MEGA11: Molecular Evolutionary Genetics Analysis version 11. *Molecular Biology and Evolution*, 38(7), 3022-3027.

Tavaré, S. (1986). Some probabilistic and statistical problems in the analysis of DNA sequences. In R. M. Miura (Ed.), *Some Mathematical Questions in Biology: DNA Sequence Analysis* (pp. 57-86). American Mathematical Society.

Xie, W., Lewis, P. O., Fan, Y., Kuo, L., & Chen, Mh. (2011). Improving Marginal Likelihood Estimation for Bayesian Phylogenetic Model Selection. *Systematic Biology*, 60(2), 150-160.

Yang, Z. (1994). Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. *Journal of Molecular Evolution*, 39, 306-314.