**Differences in phylogenetic trees of penguin taxa by FBD model and clock models**

**Introduction**

In this report, The time-calibrated trees of penguin taxa were estimated under the FBD model, strict clock model, and relaxed clock model. The data used is a morphological matrix associated with age information. The phylogenetic data derived from osteological characters consisting of 14 species of living genera, 5 fossil species, and 3 outgroups. The fossilized birth-death (FBD) model describes the stochastic processes of speciation and extinction rate. It is also used to describe the distribution of fossils and the sampling of fossil and extant species (Gavryushkina and Zhang, 2020). The idea of a molecular clock has been essential to understanding evolution for almost half a century (Zuckerkandl and Pauling, 1962). In this study, we chose two molecular clock models which are a strict clock model and a relaxed clock model. A strict clock model assumes that the evolutionary rate remains constant for every branch in a phylogenetic tree (Ferreira and Suchard, 2008). Meanwhile, a relaxed clock model assumes that each branch of a phylogenetic tree has its own evolutionary rate (Drummond et al., 2006). Since these models have the same purpose but different approaches, I proposed to see different results in penguins' time trees for different models.

**Method**

**Data**

**Morphological data**

The morphological data consist of 22 species including 14 representative species of living genera, 5 fossil species of Spheniscidae, and three outgroups. The data was derived by using 44 morphological characters. *Diomedea exulans*, *Fregata magnificens*, and *Gavia immer* were used as outgroups (Hospitaleche, 2007).

**Age information**

The data gained in this study is from the Cenozoic period. The stratigraphy maximum age is 23 Ma and the stratigraphy minimum age is 0.

**Analysis**

**tree prior (FBD) (how trees grow over time.)**

The tree model uses to describe how trees grow over time. For a tree model, the FBD model was applied. There are three parameters which are speciation rate (λ), extinction rate (μ), and fossil sampling rate (ψ) along with fossil age were used for the model. These variables were assumed as a constant rate over time and exponential prior were applied for each parameter.

These parameters are stochastic variables derived from exponential distributions with rates. We set each parameter for exponential distribution with *δ* = 10. Following, the moves were specified by setting lambda = 0.5, Tune = true for that MCMC looks for the optimum size for the move during the initial step. Subsequently, rho probability which is a constant variable was specified as 1 since extant penguins are sampled.

Next step, the origin time parameters were specified by using the age of the oldest penguin fossil as a minimum (20.0) and the age of the outgroup fossil that was used as a root species (35.0). Eventually, the moves on the tree topology and branch lengths and model were specified. The clade constraints were created to monitor specific clades as I want to focus on extant species.

The deterministic variables used in this model are turnover rate, diversification rate, and number of ancestors' samples. Due to the uncertainty of the fossil age, I used a loop to go through each dataset. To know the interval of uncertainty, the uniform distribution between minimum and maximum bounds was assigned.

**Clock models**

The clock models use to describe how evolutionary rates vary across the tree. All parameters of the tree and substitution models were set as the same since we want to focus on the results of using different clock models which are a strict clock model and a relaxed clock model.

1. **The strict clock model**

The strict clock model used to describe the rate of character changes over time assumes that the evolutionary rate remains constant for all branches in our tree. Since the rate of evolution remains unknown, an exponential prior was used for a strict clock model.

1. **The relaxed clock model**

The relaxed clock model used to describe the rate of character change over time assumes that the evolutionary rate varies for each branch and is drawn independently from the underlying distribution. The uncorrelated lognormal model was applied since it is simple and widely used. The relaxed clock model assumes each branch has an independent rate drawing from a lognormal distribution. In this model, the stochastic variables which are the mean and the variance of the lognormal distribution are defined by the estimation of the data.

Each parameter was set by using an exponential distribution with *δ* = 10. Subsequently, the moves were specified by setting lambda = 0.5, and tune = true for the optimum size for the move. Eventually, the mean of branch rates was transformed from exponential distribution to lognormal distribution.

**Substitution models**

The substitution model describes how sites evolve over time. The Mk substitution model is a generalization of the Jukes-Cantor model which assumes that the rate of change between character states is equal for all characters as well as the state frequencies. For the substitution model used, the Mk substitution model was used to describe the probability of transitioning from one character state to another.

The fnJC function was used to define a Q matrix in which our data contain 3 character states of the morphological character matrix. Furthermore, the stochastic node was defined as standard and clamped with morphological data.

**Convergence**

The effective sample size (ESS) value is the independent number drawn from the posterior distribution. If the ESS value is small, it indicates that the posterior distribution of the parameter is poor, and vice versa. The ESS value of the posterior, likelihood, and prior parameters for the strict clock model are 1206, 362, and 866, respectively. Meanwhile, the ESS value of the posterior, likelihood, and prior parameters for the relaxed clock model is 311, 401,306. For the strict clock model, the mean value for the posterior, likelihood, and prior parameters are -613.258, -505.306, and -107.952. The mean values for the posterior, likelihood, and prior parameters of the relaxed clock model are –415.013, -505.147, and 90.134, respectively. The histogram for the posterior of the strict clock model illustrates the normal distribution (figure 1A). The trace has converged between -630 to -600 (figure 2A) for the strict clock model. Simultaneously, the histogram of the relaxed clock model demonstrated right-skewed distribution (figure 1B) and the trace of the relaxed clock model converged around -520 to -220 (figure 2B). The results show that the origin time of the strict clock model is higher than the relaxed clock model, 27.09 Ma, and 26.921 Ma, respectively. The diversification rate are 9.413E-2 and 9.48E-2 for the strict clock model and the relaxed clock model. The speciation rate A screenshot of a graph

Description automatically generatedand turnover rate for both models are 0.175 and 0.44.

A screenshot of a graph

Description automatically generatedFigure 1. Posterior histogram A) Strict clock model, B) Relaxed clock model

Figure 2. Trace A) Strict clock model, B) Relaxed clock model

**Result**

According to the time-calibrated trees of penguin taxa, there is no difference between the strict clock model tree and the relaxed clock model tree in topology. In the meantime, there are slight differences in divergence time in both models. In addition, the posterior probability also varies among different lineages (figure 4). For the strict clock model tree, the divergence began at 24.92 Ma (figure 3A). Meanwhile, the first divergence appeared in 24.83 in the relaxed clock model (figure 3B). The posterior probability in both models is 1 at this node. The *Paraptenodytes antarcticus* is a sister species to an outgroup clade(f) which are *Diomedea exulans, Fregata magnificens,* and *Gavia immer,* they diverged around 16.86 Ma, and the posterior probability is 0.71 in a strict clock model and diverged at 16.58 Ma and posterior probability is 0.7 in relaxed clock model. Following, the *Palaeopheniscus* biloculata is a sister to the genus of Aptenodytes(d) and Pygoscelis(e). They diverged at 17.53 Ma and the posterior probability is 0.6 in the strict clock model and diverged at 17.49 Ma and the posterior probability is 0.62 in the relaxed clock model. Subsequently, clade c) contains *Spheniscus urbinai* and *Spheniscus megaramphus* is an ancestor group to the descendant genus which are Megadytes, Eudyptes, and the rest Spheniscus. They diverged in 17.14 Ma for a strict clock model and 16.96 Ma for a relaxed clock model. For the posterior probability, they have the same value which is 0.32 in both models. After that, *Madrynornis mirandus* diverged from different lineage around 9.48 Ma in the strict clock model and 9.47 Ma in the relaxed clock model. In this node, the posterior probability of the relaxed clocked model is higher than the strict clock model, which is 0.84 and 0.82, respectively. There was a diversification among two genera: Eudyptes(b) and Spheniscus(a), around 6.69 Ma and 6.61 Ma in the strict clock model and the relaxed clock model. At this node, the posterior probability of the strict clock model is 0.69 which is higher than the relaxed clock model, 0.68. Afterward, *Madrynornis anitipodes* is a sister to Eudyptes clade (b), they diverged around 3.91 Ma in the strict clock model and 3.95 Ma in the relaxed clock model. The posterior probability of the relaxed clocked model in this node is higher than the strict clock model, which is 0.81 and 0.77, respectively. Eventually, *Eudyptes minor* is a sister to the clade a (genus: Spheniscus). They diverged around 4.39 Ma in the strict clock model and 4.4 Ma in the relaxed clock model. According to the time-calibrated trees, the *Spheniscus humboldti* is the last descendant of the tree. It diverged from *Spheniscus magellanicus* at 0.23 Ma in the strict clock model and 0.22 Ma in the relaxed clock model. The posterior probability in the strict clock model is 0.34 which is higher than the value in the relaxed clock model, 0.32.

A screenshot of a computer screen

Description automatically generated

Figure 3. The time-calibrated trees of penguin taxa and node ages A) Strict clock model tree B) Relaxed clock model tree

**Discussion**

There is no difference in topology but divergence time in both models. In addition, the relaxed clock model tends to have lower values in divergence time compared to the strict clock model in the same branch. Considering each node in both time-calibrated trees of penguin taxa, the posterior probability in the relaxed clock model is higher than in the strict clock model. Hence, the tree from the relaxed clock model tends to be more accurate than the tree from the strict clock model since the evolutionary rates are independent for each branch.

For further study, incorporating data from several sectors would be beneficial for further study and for people who want to work on evolution and phylogenetics.

**Data availability**

https://github.com/Taeyakyk/Phylogenetics-project.git

**References**

* Drummond, A.J., Ho, S.Y.W., Phillips, M.J. & Rambaut, A. (2006). Relaxed phylogenetics and dating with confidence. PLoS Biology, 4(5), e88.
* Ferreira, M.A.R. & Suchard, M.A. (2008). Bayesian analysis of elapsed times in continuous‐time Markov chain. Canadian Journal of Statistics, 36(3), 355-368.
* Gavryushkina, A. & Zhang, C. (2020). Total-Evidence Dating and the Fossilized Birth–Death Model. The Molecular Evolutionary Clock. Springer, Cham.
* Hospitaleche, A. (2007). A new Miocene penguin from Patagonia and its phylogenetic relationships. Acta Palaeontologica Polonica, 52(2), 299-314.
* Zuckerkandl, E. & Pauling, L.B. (1962). Molecular disease, evolution, and genetic heterogeneity. Horizons in biochemistry, New York: Academic Press, 189–225.