



The Shape of Empirical Phylogenies Under Phase-Type Distributed Times to Speciation Model

Albert Ch. Soewongsono, Barbara R. Holland, Małgorzata M. O'Reilly



UNIVERSITY OF TASMANIA AUSTRALIA

albert.soewongsono@utas.edu.au

School of Natural Sciences (Mathematics)

Aims

- Develop a macroevolutionary model with general Coxian phase-type (PH) distributed waiting times to speciation.
- Derive a likelihood expression for the probability of observing any tree with branch lengths under the model with speciation but no extinction.
- Using maximum likelihood estimation and branch lengths of two empirical reconstructed phylogenies, squamates [1] and angiosperms [2], we find model parameters that maximise the likelihood.
- Derive hazard function for speciation events from the best-fitting general Coxian PH distribution to some clades in both phylogenies.

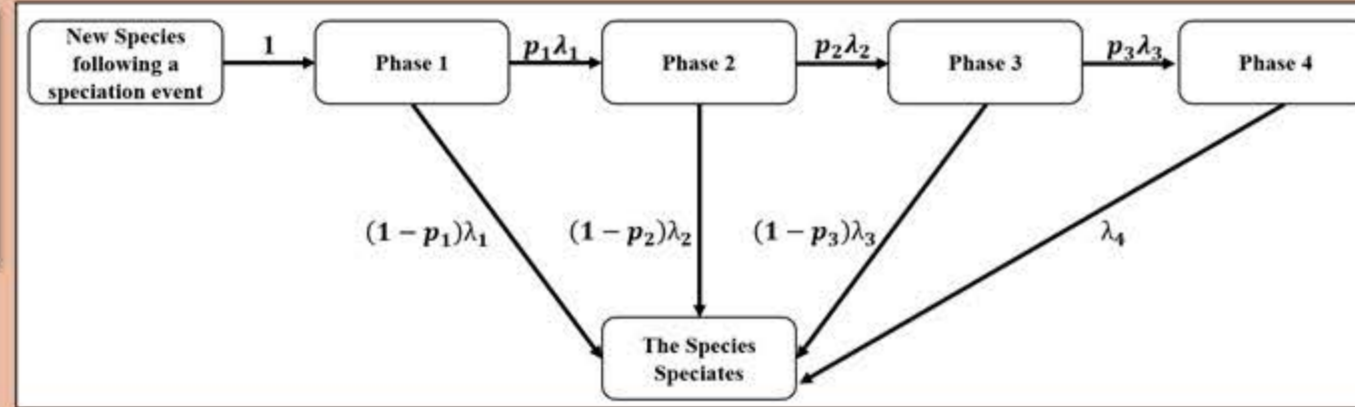
General Coxian PH Distribution

A general Coxian PH distribution $T \sim PH(\alpha, Q)$ models time T to absorption in a continuous-time Markov chain with transient states $1, \dots, n$ (referred to as phases) and an absorbing state 0.

Example:

- Initial distribution vector: $\alpha = (1, 0, 0, 0)$.
- Matrix Q of transition rates between transient states:

$$Q = \begin{bmatrix} -\lambda_1 & p_1\lambda_1 & 0 & 0 \\ 0 & -\lambda_2 & p_2\lambda_2 & 0 \\ 0 & 0 & -\lambda_3 & p_3\lambda_3 \\ 0 & 0 & 0 & -\lambda_4 \end{bmatrix}$$



$0 < p_1, p_2, p_3, p_4 \leq 1$ and state 0 corresponds to a speciation event.

Likelihood-Based Inference

- Consider a reconstructed phylogeny with 5 species evolving under a general Coxian PH distributed waiting times to speciation.
- The set of internal branches of the tree is: $\{b_1, b_2, b_3, b_4\}$.
- The set of external branches of the tree is: $\{\bar{b}_1, \bar{b}_2, \bar{b}_3, \bar{b}_4, \bar{b}_5\}$.
- Then, the probability of observing that phylogeny, assuming independence of branches, is:

$$\prod_{i=1}^4 (\alpha e^{Q b_i} \mathbf{q}) \times \prod_{i=1}^5 (\alpha e^{Q \bar{b}_i} \mathbf{1})$$

- In general, for any tree T with k internal branches and l external branches, we have:

$$P(T|\theta) = \prod_{i=1}^k (\alpha e^{Q b_i} \mathbf{q}) \times \prod_{i=1}^l (\alpha e^{Q \bar{b}_i} \mathbf{1})$$

where θ is the set of model parameters

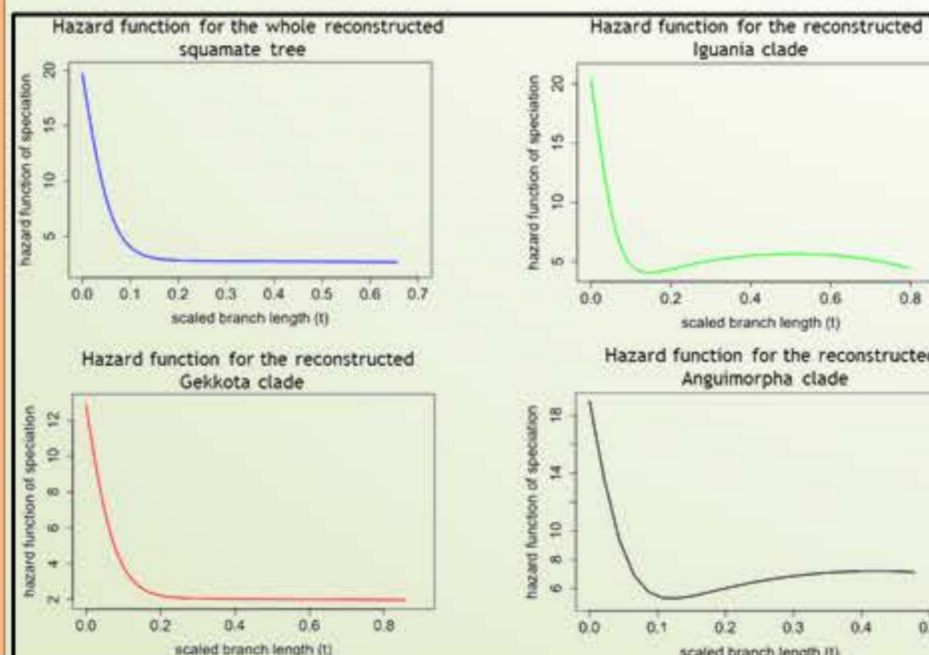
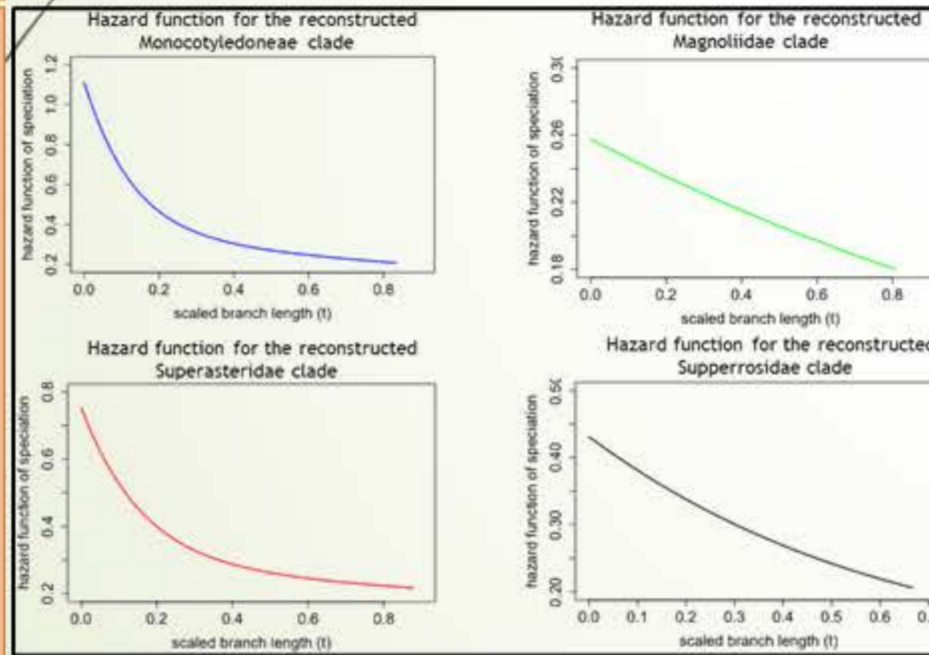
Hazard Function for Speciation

- Hazard function of speciation events is derived based best-fitting general Coxian PH distribution to each empirical clade,

$$h(t) = \frac{f(t)}{1-F(t)}$$

where t is scaled branch length, $f(t)$ and $F(t)$ are density and cumulative distribution functions, respectively.

- The hazard function describes the instantaneous rate of speciation event occurring on branches with length from a clade.
- The hazard function indicates that some clades in the squamate phylogeny experienced a shift from lower to higher speciation rates on their branches at some point.
- In general, most clades showed evidence that older species is less likely to speciate.



Model Comparison for Monocotyledoneae Clade

Model	# branches	# parameters	LogL	AIC	ΔAIC
General Coxian PH distribution	14118	7	-18439.5	36892.9	0
PH model assuming decreasing rate		3	-18788.4	37582.8	689.9
PH model assuming increasing rate		3	-45591.2	91188.3	54295.4
Exponential distribution		1	-22149.7	44301.4	7408.5
Weibull distribution		2	-18633.1	37270.1	377.2

Model Comparison for Magnoliidae Clade

Model	# branches	# parameters	LogL	AIC	ΔAIC
General Coxian PH distribution	2092	7	-3347.3	6708.6	0
PH model assuming decreasing rate		3	-3476.5	6959	250.4
PH model assuming increasing rate		3	-8925.9	17857.9	11149.3
Exponential distribution		1	-3633.4	7268.9	560.3
Weibull distribution		2	-3395.2	6794.3	85.7

Model Comparison for Superasteridae Clade

Model	# branches	# parameters	LogL	AIC	ΔAIC
General Coxian PH distribution	20016	7	-29747.6	59509.2	0
PH model assuming decreasing rate		3	-30533.9	61073.7	1564.5
PH model assuming increasing rate		3	-59551.2	119108.3	59599.1
Exponential distribution		1	-33668.5	67339.1	7829.9
Weibull distribution		2	-30064.2	60132.4	623.2

Model Comparison for Superrosidae Clade

Model	# branches	# parameters	LogL	AIC	ΔAIC
General Coxian PH distribution	11323	7	-29977.3	59968.6	0
PH model assuming decreasing rate		3	-30717.7	61441.3	1472.7
PH model assuming increasing rate		3	-72613.9	145233.8	85265.2
Exponential distribution		1	-33136.3	66274.5	6305.9
Weibull distribution		2	-30183.8	60371.7	403.1

Model Comparison for Squamate Phylogeny

Model	# branches	# parameters	LogL	AIC	ΔAIC
General Coxian PH distribution	8322	7	6503.7	-12993.4	0
PH model assuming decreasing rate		3	6466.4	-12926.7	66.7
PH model assuming increasing rate		3	2026.8	-4047.6	8945.8
Exponential distribution		1	6032.9	-12063.9	929.5
Weibull distribution		2	6213.1	-12422.2	571.2

Model Comparison for Gekkota Clade

Model	# branches	# parameters	LogL	AIC	ΔAIC
General Coxian PH distribution	1318	7	718.6	-1423.2	0
PH model assuming decreasing rate		3	714.1	-1422.3	0.9
PH model assuming increasing rate		3	255.9	-505.9	917.3
Exponential distribution		1	648	-1294	129.2
Weibull distribution		2	668.9	-1333.8	89.4

Model Comparison for Iguania Clade

Model	# branches	# parameters	LogL	AIC	ΔAIC
General Coxian PH distribution	1936	7	1598.6	-3183.1	0
PH model assuming decreasing rate		3	1587.7	-3169.4	13.7
PH model assuming increasing rate		3	489.6	-973.1	2210
Exponential distribution		1	1520.4	-3038.8	144.3
Weibull distribution		2	1538.5	-3072.9	110.2

Model Comparison for Anguimorpha Clade

Model	# branches	# parameters	LogL	AIC	ΔAIC
General Coxian PH distribution	200	7	153.3	-292.6	2.9
PH model assuming decreasing rate		3	148.8	-291.7	3.8
PH model assuming increasing rate		3	48.8	-91.6	203.9
Exponential distribution		1	146.9	-291.8	3.7
Weibull distribution		2	149.7	-295.5	0

Comparison of Different Models

- Using branch lengths from several clades in the squamate and angiosperm phylogenies, we compare different models by deriving their parameters that maximise the likelihood.
- Best-fitting model for each clade is chosen based on Akaike's Information Criterion (AIC) value [3].
- Our model provides a better fit to empirical phylogenies compared to models that follow other distributions.

References

- Pyron RA, Burbink FT, Wiens JJ (2013) A phylogeny and revised classification of squamate, including 4161 species of lizards and snakes. BMC Evol Biol 13(1):93
- Zanne AE, Tank DC, Cornwell WK, Eastman JM, Smith SA, FitzJohn RG, McGlenn DJ, O'Meara BC, Moles AT, Reich PB, et al. (2014) Three keys to the radiation of angiosperms into freezing environments. Nature 506(7486):89-92
- Akaike H (1998) Information theory and an extension of the maximum likelihood principle. In: Selected papers of Hirotugu Akaike, Springer, pp 199-213