Modeling continuous trait evolution in the absence of phylogenetic information

Mariia Koroliuk
GU university, EM in complex systems
under the supervision of Daniele Silvestro

June 29, 2014

Contents

Abstract

In this project we explore the possibility to model the evolution of continuous traits using incomplete data and in the absence of phylogenetic information. We develop a probabilistic framework to estimate evolutionary parameters under different stochastic models of trait evolution, namely Brownian motion, Brownian motion with trend and Ornstein-Uhlenbeck. This method is implemented in a maximum

likelihood and in a Bayesian frameworks. We tested the approach on simulated data to assess its performance and on and empirical data set of extant and extinct species of the dog family Caninae. We used likelihood ratio test and Akaike information criteria to find the best model.

1 Introduction

This project is devoted to the problem of a reconstructing the evolutionary history of a continuous trait in the absence phylogenetic information, using complex systems theory. Phylogenetics studies evolutionary relationships among groups of organisms, usually species or higher taxa (see [10]). The general idea is to use the information on the evolutionary relationships of organisms (phylogenetic trees) to understand the underlying process of species' history. Time of speciation and extinction for different species, trait values of the extant (present) taxa are treated as the result of a stochastic process of evolution. The project is realized using python programming language. All the graphics are made with "matplotlib" package. The last (Bayesian) part of analysis is done with a help of a program named "Tracer" [11].

1.1 Relevancy

The reconstruction of the evolution history is an important problem in modern biology. Two years ago another Erasmus Mundus project was devoted to phylogenies, in particular to a probabilistic method of estimating spetiation and extinction rates. More details can be found here [4] and here [5] .

In this project we propose a new method that allows us to fit stochastic models of continuous trait evolution using exclusively fossil information without attempting to reconstruct the phylogeny but instead accounting for all possible underlying phylogenetic relationships. This feature is extremely useful, when a phylogenetic tree can not be reliably reconstructed,which is very common in many data sets based on fossil records, such as the Canidae data set discussed below. This could be a problem, because accurate phylogenetic trees require DNA information, which is available only from recent history, while many species got extinct.

Here we implement three of the most commonly used models of stochastic trait evolution, i.e. Brownian motion (BM), Brownian motion with trend (BMT), and Ornstein-Uhlenbeck (OU) models [2], using both maximum likelihood and Bayesian frameworks to compute their relative fit and estimate the parameters. Then we will test this method by simulations as well as on empirical data set from the dog family, Canidae. We consider the case with one trait value for each species and assume that the trait evolution will occur at the speciation events, whereas the trait is assumed to be constant throughout the lifespan of a species. The main aims of this study are:

1)to compute the likelihood of different models of trait evolution without using phologenetic information

2)to estimate the parameter values under BM, BMT and OU models

3)to select the best fitting model

2 Methods

2.1 Notations

Let us consider a set of N extant and extinct species for which fossil record is available documenting their time of origination and extinction ($\mathbf{t} = \{t_1, ..., t_N\}$, $\mathbf{e} = \{e_1, ..., e_N\}$) and the value of a continuous trait ($\mathbf{v} = \{v_1, ..., v_N\}$). In our notation, the ages of all events are measured as the time before the present and the species are sorted by their origination time so that $t_1 > t_2 ... > t_N$.

2.2 Likelihood functions

2.2.1 Theory and definition

Brownian Motion, or BM is is a centered Gaussian process $\{W_t\}_{t\geq 0}$ with a covariance function $R(t, s) = \mathsf{E}[W_t W_s] = \sigma^2 \min(t, s)$.

Obviously, this definition implies that $W_0 = 0$ a.s. and the process $W_{tt\geq 0}$ has stationary, independent increments. One can also obtain that $t \to W_t$ is continuous a.s. We emphasize that BM is characterized by one rate parameter σ^2.

Brownian Motion with trend, or BMT with a trend μ is a process $\{\hat{W}_t\}_{t\geq 0}$ defined as $\hat{W}_t = W_t + \mu_0$, and thus characterized by the parameters σ^2 and μ_0. BMT adds a temporal variation of the expected mean trait value (according to a linear variation with slope μ_0). BMT with a trend $\mu = 0$ is simply a BM.

Ornstein-Uhlenbeck process is a centered Gaussian process $\{X_t\}_{t\geq 0}$, such that $\mathsf{cov} X_t X_s = \rho e^{-\alpha|t-s|}$ and $X_0 \sim \mathcal{N}(0, \rho)$.

Let X be a random variable with a discrete probability distribution P depending on a parameter θ.

The likelihood of a set of parameter values, θ , given outcomes x, is equal to the probability of those observed outcomes given those parameter values, that is $\mathcal{L}(\theta|x) = P(x|\theta)$.

In case of a discrete probability distribution likelihood function is $\mathcal{L}(\theta|x) = p_\theta(x) = P_\theta(X = x)$, considered as a function of θ.

For continuous probability distribution: If a random variable possesses a density f of a distribution depending on a parameter θ. Then the function $\mathcal{L}(\theta|x) = f_\theta(x)$ considered as a function of θ, is called the **likelihood function**. We would use this definition below.

For example, for BM with a rate parameter σ^2:

$$P_{BM}(x|\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{x}{2\sigma^2}\right) \tag{1}$$

For BMT with a rate parameter σ^2 and trend μ_0:

$$P_{BMT}(x|\sigma^2, \mu_0) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-\Big[x - \mu_0\Big]^2}{2\sigma^2}\right] \tag{2}$$

OU combines the stochastic component of BM with a selection parameter (α) that leads trait values to and maintains them around an optimum μ_0. The likelihood function is as the following:

$$P_{OU}(x|\sigma^2,\alpha,\theta,y) = \sqrt{\frac{\theta}{\pi\sigma^2[1-e^{-2\theta}]}}\exp\left[\frac{-\theta\Big[x-ye^{-\theta}\Big]^2}{\sigma^2\Big[1-e^{-2\theta}\Big]}\right] \tag{3}$$

Since phylogenetic information is considered to be unavailable, we need to consider all possible evolutionary links between any pair of species, the only constraint being the origination times of the species. Hence, a species trait value v_j at time t_j can only derive from species that originated prior to time t_j, i.e. $t_1,...t_{j-1}$ and all of such species can potentially be the parent lineage. The amount of evolutionary change in the trait values between a parent species i and its descendant j is calculated by $v_j - v_i$ and indicated with x in our notation. In terms of definition of the likelihood,observed data is v_i, σ^2 for BM , v_i, σ^2, μ_0 for BMT, $v_i, \sigma^2, \alpha, \theta$ for OU, and outcomes x is v_j at all cases. So likelihood functions can be rewritten as:

$$P_{BM}(v_j|v_i,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(\frac{-(v_i-v_j)^2}{2}\right) \tag{4}$$

$$P_{BMT}(v_j|v_i,\sigma^2,\mu_0) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[\frac{-\Big[(v_i-v_j)-\mu_0\Big]^2}{2}\right] \tag{5}$$

$$P_{OU}(v_j|v_i,\sigma^2,\alpha,\theta) = \sqrt{\frac{\theta}{\pi\sigma^2[1-e^{-2\theta}]}}\exp\left[\frac{-\theta\Big[v_j-\alpha-(v_i-\alpha)e^{-\theta}\Big]^2}{\sigma^2\Big[1-e^{-2\theta}\Big]}\right] \tag{6}$$

Hereafter, we will use the notation $L_{ij} = P(\Delta(v_{ij})|\Theta)$ where Θ represents a set of parameters and $\Delta(v_{ij})$ is the evolutionary change between v_i and v_j.

2.2.2 Application to the algorithm

Likelihood of a data set j to have evolved to certain trait value v_j conditional on all possible links to older species, i.e. accounting for all possible evolutionary paths, is given by:

$$L_j = \sum_{i=1}^{j-1}(L_{ij}), \tag{7}$$

which can be extended to the likelihood of the entire data set , accounting for all unknown evolutionary links:

$$P(\mathbf{v}|\mathbf{t},\varphi) = \prod_{j=2}^{N}\left[\sum_{i=1}^{j-1}(L_{ij})\right] \tag{8}$$

All likelihood values in this report are given in log form as the product in equation (8) generates numbers that are typically too small to be tractable.

The actual value of a likelihood function bears no meaning, however it can be used to optimize the parameters values and to compare the relative fit of different models (BM,BMT and OU) it's use lies in comparing one value with another. In part one, algorithm would maximize values of likelihood for different possible sets of parameter and take the maximal one. In part two we will use an MCMC algorithm to find the posterior estimate of the parameters along with their 95% credible intervals.

2.3 MCMC

In general, MCMC algorithm could be described as:

1. Start at random initial parameter values.

2. Make a small random move.

3. Calculate the posterior ratio between the current state and the previous state, here indicated with r.

4.If $r > 1$ always accept the current parameter values, if $r < 1$ accept the current parameter values with probability equal to r. Accepted values are stored as posterior samples and will be used to compute mean and 95% credible intervals.

5.Go back to state 2.

In our case step 2 works as follows: firstly, we update parameter values by a small random number. Then new likelihood, prior and posterior are calculated. To calculate the prior probabilities of the parameters we use a gamma distribution for the parameters that must take positive values (such as σ and α) and normal distribution centered in 0 for parameters that can take both negative and positive values (e.g. μ_0).

For models that have several parameters,as BMT and OU, at every MCMC iteration we update one parameter chosen as random.

In the figure 1 we can see an example of a run made by Tracer program. At the beginning, there is a period of dramatical growth, typically named burn-in phase (highlighted in gray color here), which would be ignored afterwards at the analysis. Then it stars fluctuating, and this fluctuations are represented in a histogram 1. Red lines represents 95% value appearance interval , which is named credible interval further.

2.4 Simulation

Times of origination and extinction are simulated according the birth-death process as in [1]. Afterwards values of trait are simulated along with the genetic three history according the following equations:

$v(i) = 1$, if i is the oldest species species in the data set.

If i_1 is a parent of i_2 then:

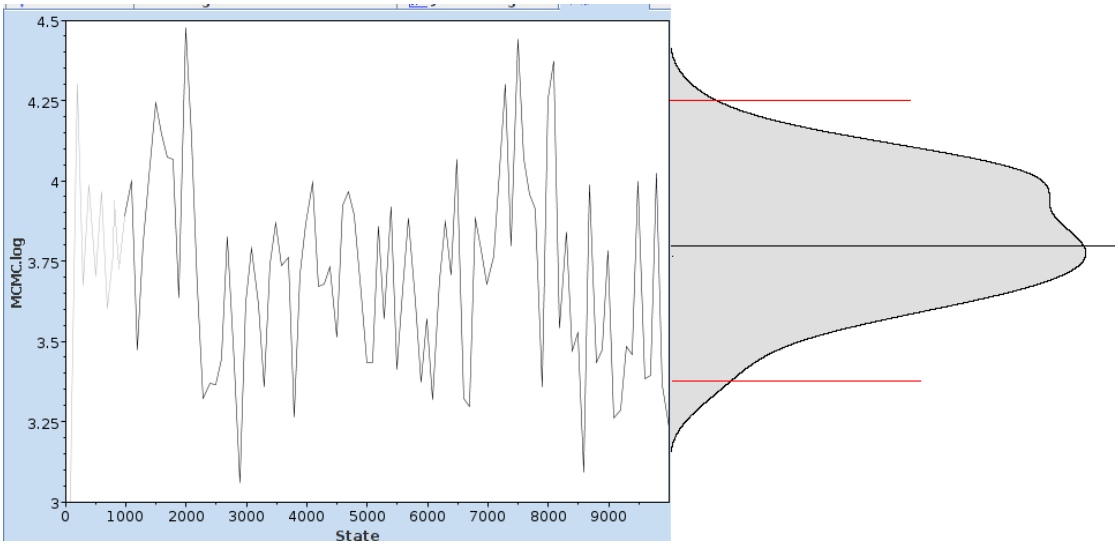


Figure 1: example of a run. MCMC

$$v(i_2) = v(i_1) + \mathcal{N}(0, \sigma),$$

in case of BM with the rate parameter σ^2

$$v(i_2) = v(i_1) + \mathcal{N}(\mu_0 \Delta t, \sigma),$$

in case of BM with the rate parameter σ^2 and trend mu_0

$$v(i_2) = \mathcal{N}((v(i_1)exp(-\theta) + \mu_0(1 - exp(-\theta)),$$

$$\sigma\sqrt{\frac{1 - \exp(-\theta)}{2\theta}})$$

in case of OU with the parameters σ^2, θ and mu_0.

2.5 Methods for model definition

It is important to chose the best fitting model out of 3 possibilities. Here we use Akaike Information Criterion and Likelihood Ratio Test to quantify the relative model fit of BM, BMT, and OU.

The Akaike Information Criterion (AIC) is a way of selecting a model from a set of models. The chosen model is the one that minimizes the Kullback-Leibler distance between the theoretical and experimental model. It is based on the information theory. A heuristic way is to consider this criterion as one that seeks a model that has a good fit to the truth but few parameters. It is defined as

$$AIC = -2(\ln(L)) + 2K$$

where L is the probability of the data conditioned on a model and K is the number of free parameters in the model.

Kullback-Leibler distance for 2 absolutely continuous probability measures P and Q is defined as

$$D_{\mathrm{KL}}(P\|Q) = \int_{-\infty}^{\infty} \ln\left(\frac{p(x)}{q(x)}\right) p(x)\,\mathrm{d}x$$

where p and q denote respectively the densities of P and Q.(see [7],[6] for more information)

Likelihood ratio test, or LRT, is used to compare the fit of two models, one of which is nested within the other. The test statistic (usually denoted D) is : $D = -2ln(L_0) + 2ln(L_1)$, where L_1 is the likelihood of an alternative model and L_0 is the likelihood of the null model. The value of D then is compared with the value of χ^2 distribution with degrees of freedom equal to $df_1 - df_0$. Symbols df_0 and df_1 represent the number of free parameters of the null model and the alternative model. In this analysis we will use χ^2 distribution at 95% confident level.

3 Results

3.1 Calculations of errors

The error (or disturbance) of an observed value is the deviation of the observed value from the (unobservable) true value. So if $\phi = \widehat{\eta}$ is a stochastic variable that corresponds to the estimation recover parameter , then $\phi = \eta + \epsilon$, where ϵ is an error. The idea of this subsection is to validate the algorithm by investigating the value of biases. To test we simulated the data under known parameter values, optimized the parameters using maximum likelihood, and computed the error of the estimates. This was repeated after removing at random different fractions of the species in the data set to reproduce the case of incomplete sampling. The fraction of species left in the data set is indicated by ρ. For example, a value of $\rho = 0.3$ means that we removed 70% of data.

3.1.1 Parameter estimations bias dependent on removed data

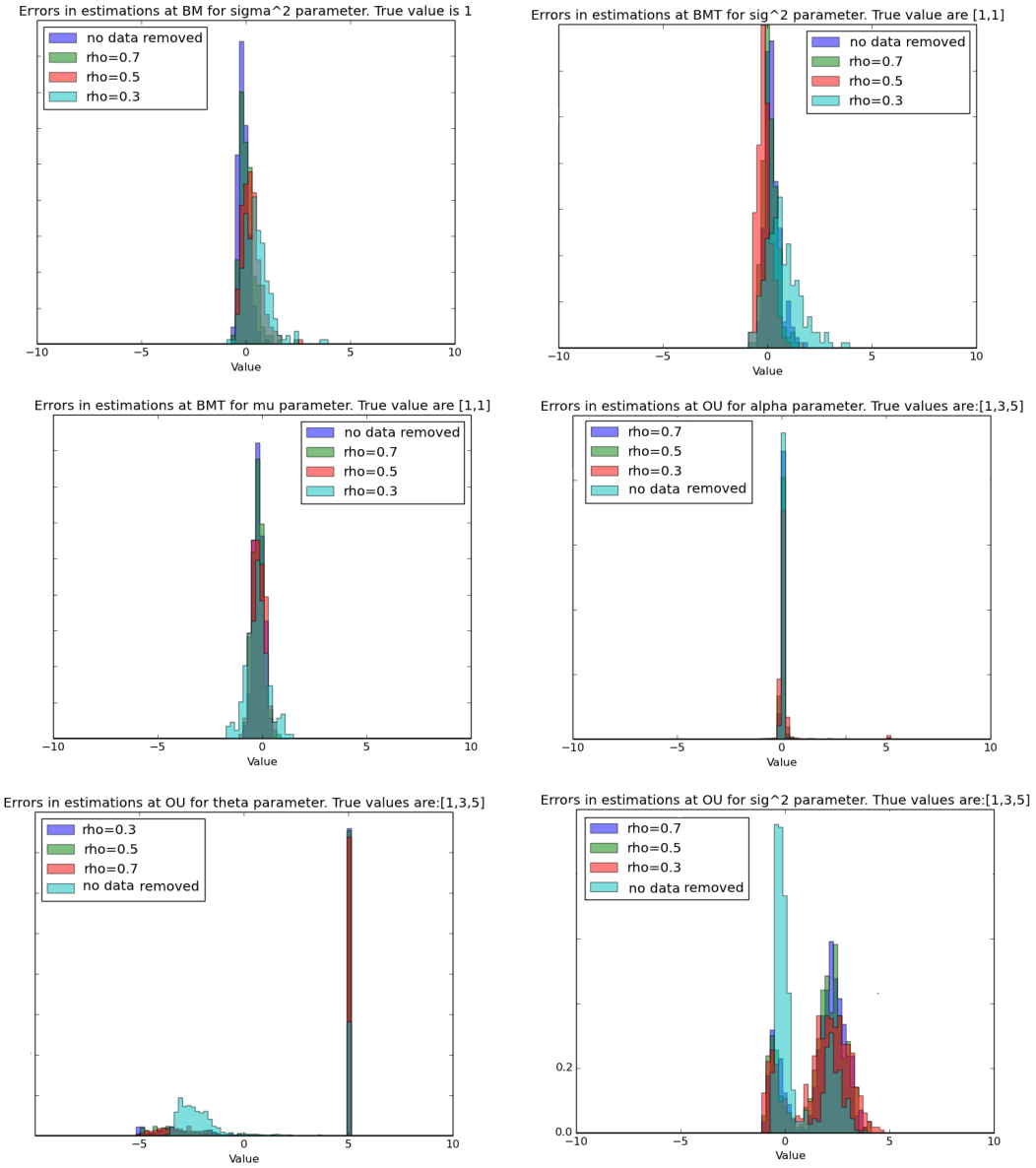
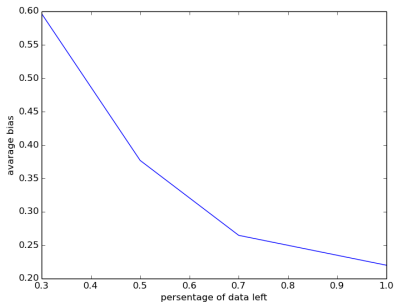


Figure 2: Parameter estimations bias

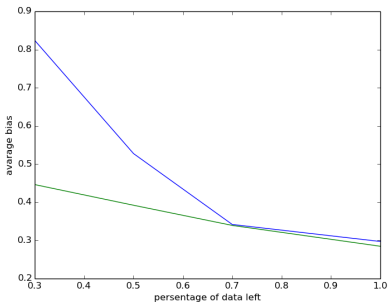
Figure 3.1.1 shows histograms of errors out of 1000 runs of the simulation, thus, for example, we can see maximal and minimal error, mean error, etc. For different values of ρ different colors are used.

Figure 3.1.1 shows average value of a bias depends on amount of removed data, ρ.

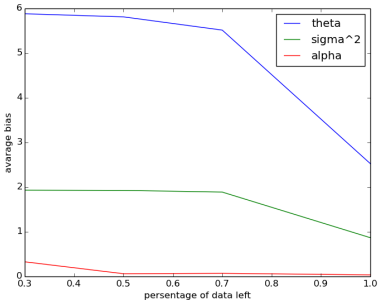
There are some observations we can formulate from these plots:



(a)



(b)



(c)

Figure 3: comparison of average biases dependent on a value of removed data ρ

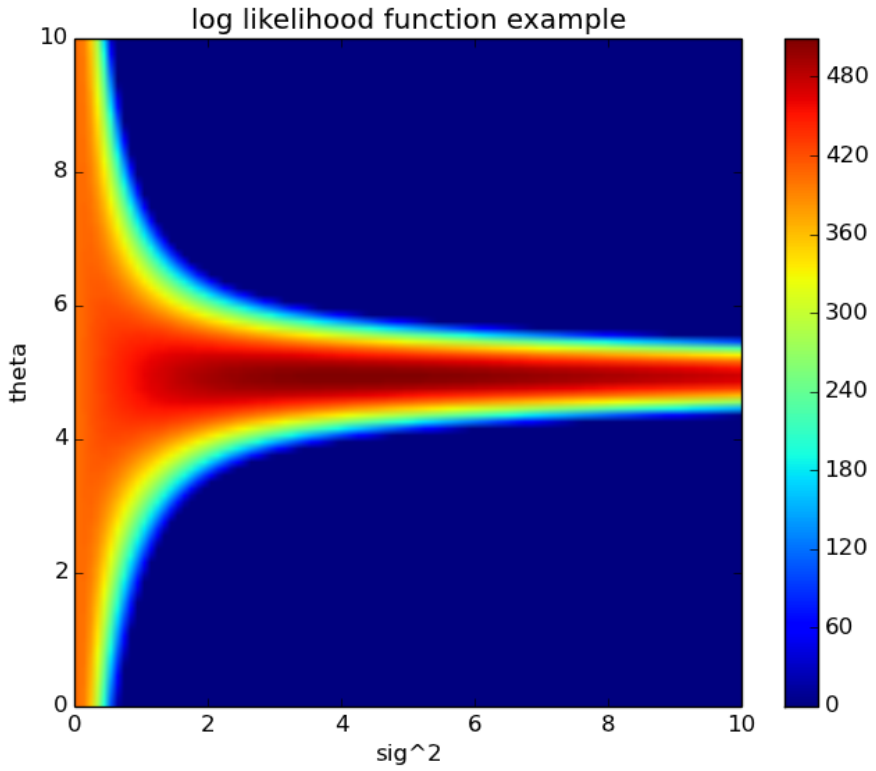
1)The behavior of the σ^2 parameter bias in case of BM and BMT acts the same, which is understandable as BMT is a generalization of BM, that saves the meaning of σ^2.

2)With decreasing values of ρ, the bias tends to increase towards positive values. This happens because links between ancestral species and their direct descendants disappear from the data, thus increasing the amount of change in trait values between the remaining species.

3) Estimation residuals for μ_0 parameter for BMT and α for OU is centered around 0, indicating no trends towards positive or negative biases. These are features of a good estimation. To prove this hypothesis more precisely that graphical analysis we can use χ^2 test. We will test at 5% significance level two hypotheses: $H_1, 0$ - expectation of error is zero, $H_2, 0$-distribution of error is normal. Results of this test show that the error is unbiased and normally distributed.

4) One of the most unexpected behavior is shown by σ^2 of OU. It shows positive bias even if data is not removed. One more difference is that decreasing of ρ (less data saved) changes the mean error logarithmically.

5) Estimation of θ value shows huge positive bias. The peak at +5 is created artificially because estimated value of θ is limited by 10, so +5 shows when it reached the

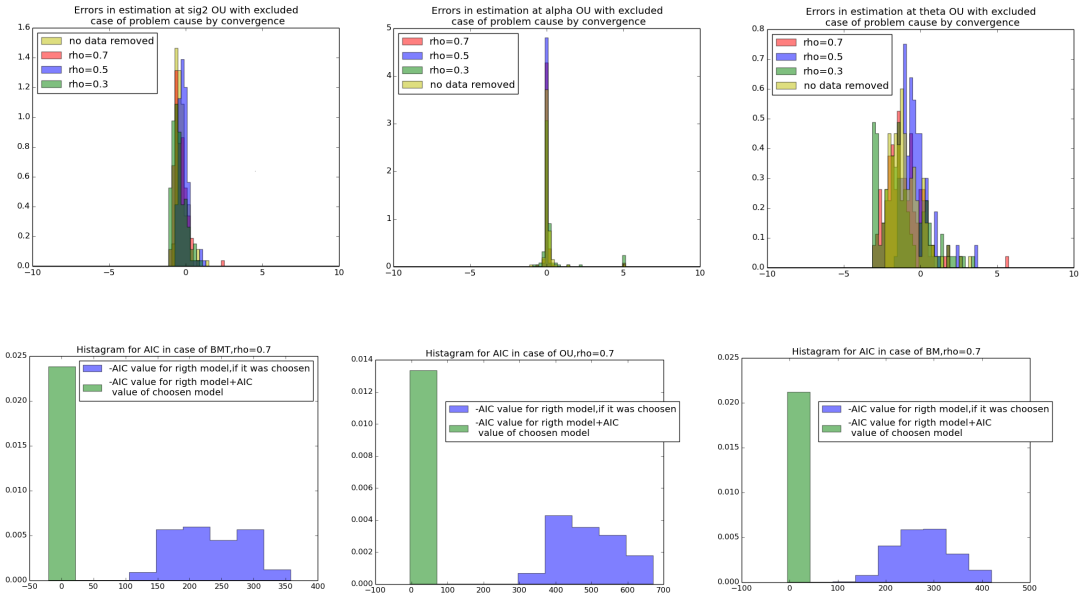


boundary. Worth to notice, that mean error estimation for θ remains relatively consistent with a change of ρ. The problem with overestimating of θ is caused by an convergence to an optimum problem. It means that search is stuck at the edge of allowed area. In the figure 3.1.1 we can see an example of the run where the problem with convergence did happen. In this image values of log likelihood function are plotted as a function of θ and σ^2. Note, that only positive values are plotted. Actual maximum(according to the simulated values) is supposed to be at ($\sigma^2 = 2$, $\theta = 5$). In fact, there is a region of increased log likelihood at that point. However, found result is ($\sigma^2 = 3.25$, $\theta = 10$). Corresponding likelihood values are almost the same hight, so finding the real maximum could fail (as it occurred in this example).

To conclude, the convergence problem is a big issue in this estimation that might require improved likelihood optimizers to be solved. If do not take into account cases when that problem did occurred, histograms transforms to 3.1.1 , which shows better results comparing to 3.1.1.

3.2 Model selection

In this section we will use AIC and LRT to determine which model fits better.



Results of LRT are shown as table, separated for two nested pairs of hypothesis was tested- BM against BMT and BM against OU. Only part of "false positive" (when preferred model is simpler that simulated one) and "false negative"(when preferred model parameter rich that the simulated one) are specified.

	BM-BMT		BM-OU	
	BMT (true BM)	BM (true BMT)	BM(true OU)	OU (true BM)
$\rho = 0.3$	48%	22.2%	39.5%	44.2%
$\rho = 0.5$	30.7%	11.2%	35.6%	21.9%
$\rho = 0.7$	19.6%	6.9%	28%	12.1%
$\rho = 1$(no data removed)	4.4%	6.1%	22%	1%

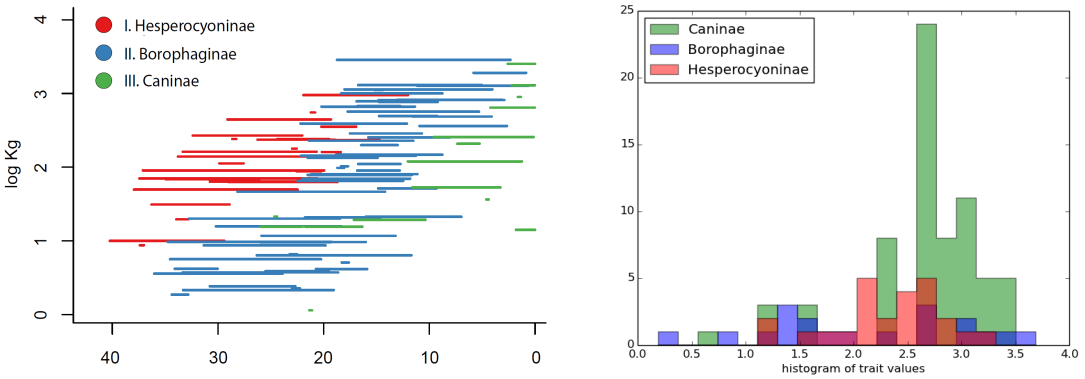
Table 1: LRT results

In general, we find the correct model with hight accuracy when the data are complete, but model selection becomes more difficult as more species are removed from the data.

4 Application

In this section we will apply the methods described above to analyze fossil data from species of the family Canidae, carnivores that includes domestic dogs, wolves, foxes,

jackals, coyotes, and many other lesser known extant and extinct dog-like mammals. We analyzed the evolution of their body mass (lo transformed for the analysis) and the times of extinction and spetiation are obtained from the analysis of fossil occurrence data using the program PyRate [12]. There are three families represented at the data set: Hesperocyoninae (29 species), Borophaginae (68 species) and Caninae (23 species). In total data set have 120 species.

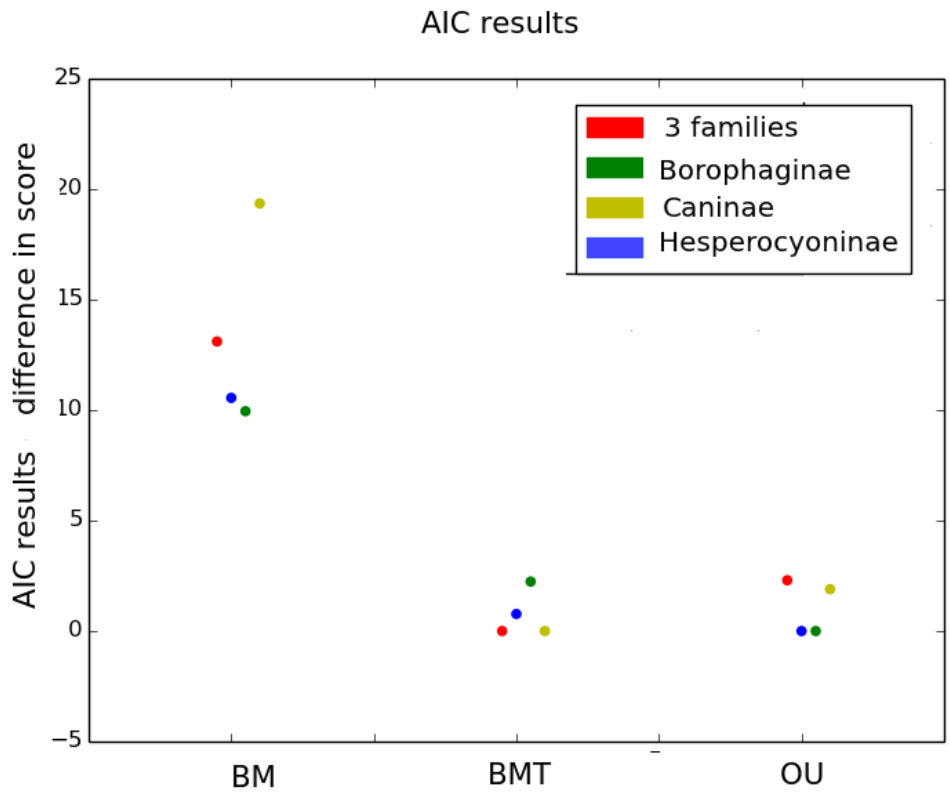


(a) (b)

In the figure 4 we can see representation of this data in a graphical format,histogram of trait values for different families and species diversity through the time line.
In the next table L1,L2,L3 corresponds to the log likelihood of all three hypothesis (BM,BMT and OU respectively). AIC means the result of AIC test, LRT1 refer to the chosen model out of pair BM-BMT, LRT2 refer to a chosen model out of BM-OU.

	BM:L_1	σ^2	BMT:L_2	σ^2	μ_0	OU:L_3	σ^2	α	θ	AIC	LRT1	LRT2
3 fam.	257.087	0.162	264.639	0.127	0.255	264.49	0.144	10	0.032	BMT	BMT	OU
Hesperocyoninae	35.422	0.146	41.313	0.058	0.33	42.699	0.108	3.247	0.331	OU	BMT	OU
Caninae	12.217	0.471	17.072	0.171	0.633	19.192	0.088	5.174	0.23	OU	BMT	OU
Borophaginae	111.853	0.195	122.529	0.187	0.428	122.583	0.219	10	0.053	BMT	BMT	OU

In the graph 4 we can see values of AIC plotted for all three models for every family.



The weakest model always is BM, while BMT and OU obtain similar support with OU being the best for Hesperocyoninae and Caninae family and BMT for Borophaginae and the three subfamilies as a whole. Noticeably, the models where α parameter was estimated as 10 are not chosen,so they are not the best. This can be an evidence that evolution of body size in Canidae does not follow an unconstrained random process (as the BM), but showed temporal trends (BMT),potentially towards optimal values (OU). Optimal value for OU shows log of the optimal mass ,so for example, in case of Hesperocyoninae family optimal value was around 20 kg, which is bigger that observed mean . Also as we can see, value for θ for BMT are positive, so there is a tendency of increasing values. Observation from the data 4 prove this point (we can see shift to the up direction over the time).

In the following section we analyze the Canidae data set using a Bayesian implementation of them model, from which we can derive,not only the most probable values, but also it's credible intervals. The number of iteration was 25000 ,and the sampled parameter values were saved to a file every 100th iteration for further analysis. Results of MCMC and its comparison with maximum likelihood are presented at the following tables.

Family	σ^2	mean of σ^2 with MCMC	95% HPD interval MCMC
3 families	0.145	0.174	[0.1064,0.2454]
Hesperocyoninae	0.146	0.1554	[0.0527,0.2913]
Caninae	0.471	0.4229	[0.1238,0.8646]
Borophaginae	0.074	0.052	[0.0336,0.075]

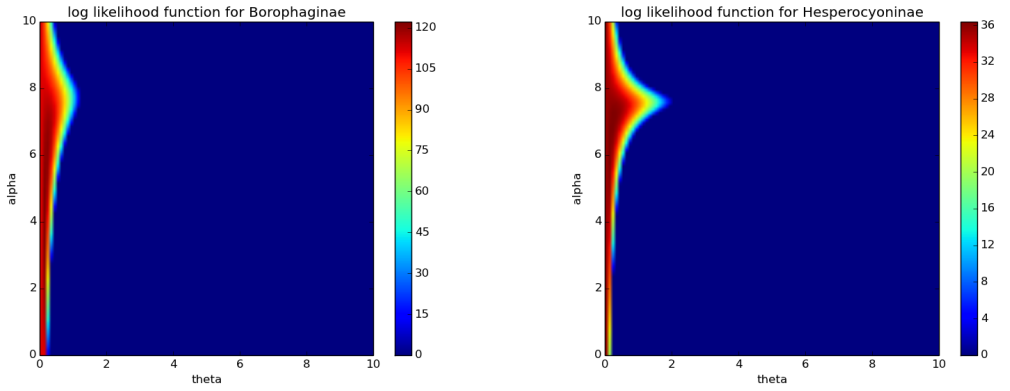
Table 2: results for BM ,comparison of method of max likelihood and MCMC.

Family	σ^2	mean of σ^2	95% HPD	μ_0	μ_0	95% HPD
3 families	0.127	0.159	[0.0606,0.263]	0.255	0.2671	[0.09055,0.3948]
Hesperocyoninae	0.187	0.1002	[0.0293,0.2161]	0.428	0.3337	[0.1229,0.5317]
Caninae	0.172	0.3	[0.0677,0.7266]	0.633	0.6505	[0.2871,1.0626]
Borophaginae	0.187	0.2315	[0.0935,0.4081]	0.428	0.4509	[0.2371,0.69]

Table 3: results for BMT,comparison of method of max likelihood and MCMC.

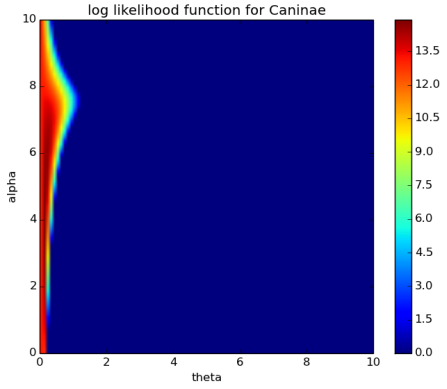
Family	σ_2	$\sigma_2 MCMC$	95%HPD	α	α MCMC	95% HPD	θ	θ MCMC	95% HPD
3 families	0.144	0.3756	[0.07,0.76]	10	5.0194	[2.1382,8.9]	0.032	0.1938	[0,0.4416]
Hesperocyoninae	0.108	1.9	[0.053,5.26]	3.2	2.6	[2.1,3.53]	0.3	3.7	[0.16,9.46]
Caninae	0.088	4.5	[0.08,9.4]	5.174	2.708	[1.6508,4.6746]	0.23	2.71	[0.1146,6.3828]
Borophaginae	0.219	0.34	[0.103,0.6539]	10	6.28	[3.32,10]	0.053	0.1318	[0.03,0.32]

Table 4: results for OU,comparison of method of max likelihood and MCMC.

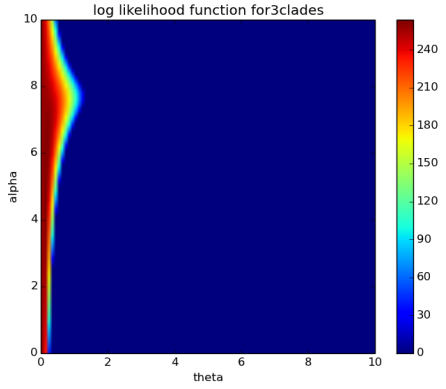


(a)

(b)



(c)



(d)

Figure 4: log likelihood surface for different families (OU)

The last part is to compare both approaches, which will be done at example of OU, as the most paramter-rich model. At 4 we can see images of likelihood surface as function of α and θ (In figure 3 the color represents the value of likelihood according to the legend at right hand side). It is noted that only points with positive values of likelihood are plotted. We can see, that found by MCMC values and intervals are situated at the area of the hight probability, so we can say that both approaches agree.

5 Conslusion

With this work we showed that trait evolution can be reconstructed even in the absence of phylogenetic information. Although we observe some limitations in the accuracy of the parameter estimates, this framework represents a step forward in the analysis of important evolutionary processes, such as body mass evolution, using fossil data. In the

case of the dog family Canidae, our analyses revealed that trends towards larger body sizes can be inferred from the data, which might have interesting biological implications. Future work should focus on improving the accuracy of parameter estimation, potentially by explicitly accounting for the incompleteness of the data and on introducing additional and more complex models of trait evolution.

References

[1] Daniele Silvestro, Jan Schnitzler,Lee Hsiang Lion ,Alexandre Antonelli, Nicolas Salamin(2014).Bayesian Estimation of Speciation and Extinction from Incomplete Fossil Occurrence Data. Syst. Biol. 0(0):1–19, 2014.

[2] Brian C. O'Meara,Evolutionary Inferences from Phylogenies: A Review of Methods,Department of Ecology and Evolutionary Biology, Annu. Rev. Ecol. Evol. Syst. 2012. 43:267–85.

[3] A Bayesian framework to estimate diversification rates and their variation through time and space,Daniele Silvestro, Jan Schnitzler, Georg Zizka,2011,BMC evolutionary biology,Volume 11,Issue 1

[4] Felsenstein, J. 1985. Phylogenies and the comparative method. American Naturalist 125:1-15.

[5] Harvey, P. H., and M. D. Pagel. 1991. The comparative method in evolutionary biology. Oxford University Press, Oxford. 239 pp.

[6] Jiusun Zeng, Lei Xie, Uwe Kruger, Jie Yu, Jingjing Sha, Xuyi Fu,Process Monitoring based on Kullback Leibler, Divergence, European Control Conference (ECC),July 17-19, 2013, Zürich, Switzerland.

[7] Akaike H.1973 Information theory and extension of the max likelihood principle, Akademiai KIado, Budapest.

[8] A Gentle Introduction toMarkov Chain Monte Carlo (MCMC),Ed Georg,University of Pennsylvania,Seminaire de Printemps Villars-sur-Ollon, Switzerland ,March 2005

[9] Phylogenetic Comparative Analysis: A Modeling Approach for Adaptive Evolution Marguerite. A. Butler ,Aaron A. King .Department of Ecology and Evolutionary Biology, University of Tennessee, Knoxville, Tennessee 37996-1610

[10] Felsenstein, J.2004 . Inferring Phylogenies. Sinauer Associates, Sunderland,Mass.

[11] Rambaut & Drummond,2009

[12] Silvestrov et al. 2014