

Introduction to Phylogenetics



Part Two

ETH zürich

02.12.21

Bethany Allen

Computational Evolution, D-BSSE

What is phylogenetics?

The science of inferring
evolutionary relationships,
often visualised using a phylogeny
or “tree”

02.12.21

Phylogenetics workshop – Bethany Allen

Slide 2 of 30

At the end of the last session I mentioned that there are three main ways of evaluating phylogenies, and described two of them. I'll now introduce the third.

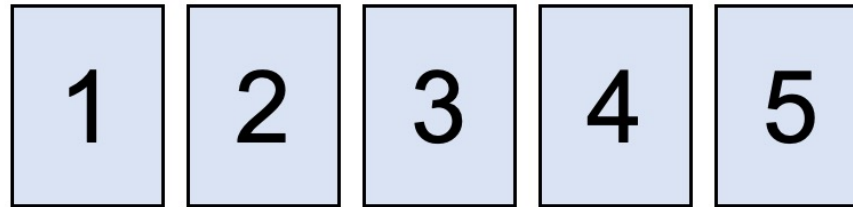
What is Bayesian phylogenetics?

Phylogenetics based on the implementation of Bayes' Theorem

Bayes' Theorem considers that the probability of an outcome is influenced by prior knowledge

This description might seem a bit confusing but we often use Bayesian thinking in day-to-day decision making.

What is Bayesian phylogenetics?



Which card will your friend choose?

Which card will your friend choose, knowing their favourite number is 3?

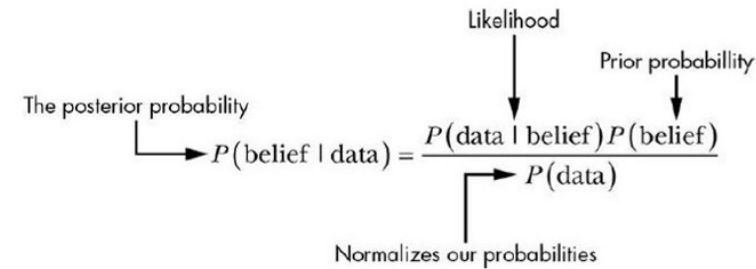
02.12.21

Phylogenetics workshop – Bethany Allen

Slide 4 of 30

Imagine you attend a fete or fair with a friend, and there is a game where you pay a franc and receive a mystery prize. The prize is determined by the number of the card you choose. Your friend decides to play, so what card do you think they will choose? In a frequentist mindset, you might think that there is a 20% chance of them choosing each of the five cards. However, you know that your friend's favourite number is three. This means that instead you can apply a Bayesian mindset and conclude that your friend will probably choose three, with say 80% likelihood, and 5% for each of the rest. This is a really simple example of how you can apply prior knowledge to inform your probabilities in a given scenario.

What is Bayesian phylogenetics?



The diagram illustrates the mathematical expression of Bayes' Theorem for phylogenetics. It features the equation $P(\text{belief} | \text{data}) = \frac{P(\text{data} | \text{belief}) P(\text{belief})}{P(\text{data})}$. Annotations include: 'Likelihood' with an arrow pointing to $P(\text{data} | \text{belief})$; 'Prior probability' with an arrow pointing to $P(\text{belief})$; 'The posterior probability' with an arrow pointing to $P(\text{belief} | \text{data})$; and 'Normalizes our probabilities' with an arrow pointing to the denominator $P(\text{data})$.

$$\text{The posterior probability} \rightarrow P(\text{belief} | \text{data}) = \frac{\overset{\text{Likelihood}}{P(\text{data} | \text{belief})} \overset{\text{Prior probability}}{P(\text{belief})}}{\underset{\text{Normalizes our probabilities}}{P(\text{data})}}$$

Shashank Parameswaran

02.12.21

Phylogenetics workshop – Bethany Allen

Slide 5 of 30

This is how we mathematically express Bayes' Theory. We need a likelihood of the data given our belief, a probability based on our belief, known as the prior, and a general probability of the data. This then gives us an overall probability, known as the posterior.

What is Bayesian phylogenetics?

The priors allow the incorporation of uncertainty in the information added to a model

The posterior provides uncertainty on the probability of an outcome, **if the priors are correct**

Our previous knowledge, the priors, are expressed in the model as distributions. That means that we can provide a range, indicating things that we think are more or less likely within that range based on what we know. In return, we also get our output probability, the posterior, as a distribution, which captures uncertainty. This can be incredibly useful. However, the analyses are fairly strongly attached to the priors, so making false assumptions can be problematic. As a result, setting priors which are a bit more general and vague can be a good idea, just in case your preconception isn't quite right.

What is Bayesian phylogenetics?

- 1) Build a model of evolution
- 2) Put priors on parameters in the model
- 3) Iterate across different prior states/values
- 4) Evaluate the fit of the output against the data
- 5) Estimate probabilities for different prior states/values

02.12.21

Phylogenetics workshop – Bethany Allen

Slide 7 of 30

This is a general pipeline for carrying out Bayesian phylogenetic analyses. We build (or choose) a model of evolution, research and apply priors, use an iterative process to select values from our priors, test how well those values fit the data, and bring all of that information together to assess what the most probable phylogeny is. I will now walk you through an example of a Bayesian analysis.

The fossilised birth-death process

Journal of Theoretical Biology 267 (2010) 396–404



Sampling-through-time in birth–death trees

Tanja Stadler*

Institut für Integrative Biologie, ETH Zürich, Universitätsstr. 16, 8092 Zürich, Switzerland

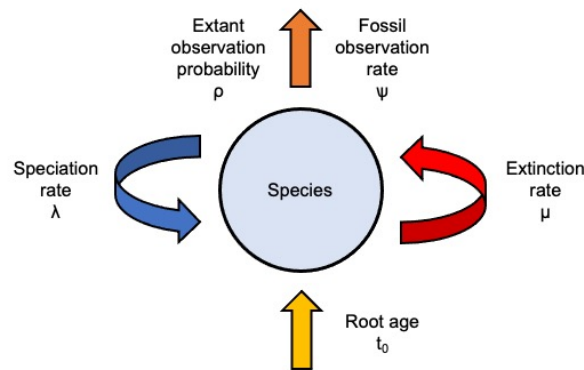
02.12.21

Phylogenetics workshop – Bethany Allen

Slide 8 of 30

In my research I will be using the fossilised birth-death process. In part this is because it was developed by my boss, Tanja Stadler, but also this is because it's a great model which seems to perform well, especially if you care about fossils, which I do. The model was outlined initially in this paper but has since been developed to broaden its applicability. I'll be describing it here in its simplest form to try and keep things understandable!

The fossilised birth-death process



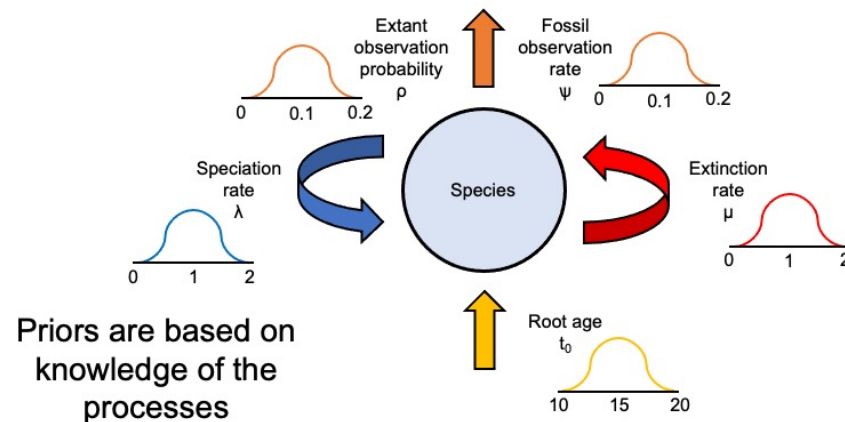
02.12.21

Phylogenetics workshop – Bethany Allen

Slide 9 of 30

The model considers a single lineage which represents a species. That lineage has a speciation rate, which is the frequency with which another lineage splits off from it. Its first speciation event occurs at a particular time, which is the root age of the clade. It also has an extinction rate, which is the frequency with which it ceases to exist. Along that lineage, fossils are preserved with another given frequency. These processes continue across all of the lineages in the phylogeny for a particular amount of time (relative to the root), at which we reach the present, and the number of lineages still existing at that point are sampled with a given probability. These five parameters describe the model.

The fossilised birth-death process



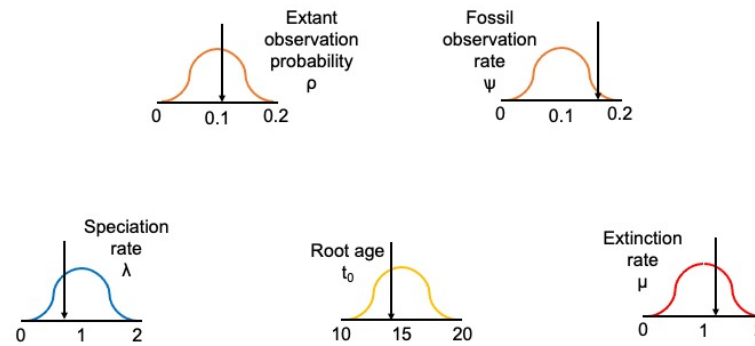
02.12.21

Phylogenetics workshop – Bethany Allen

Slide 10 of 30

To apply it, we put prior distributions on each of these parameters. It's not an easy process, and often involves searching the literature for suggestions others have made. In some ways, this is the most difficult part of the process.

The fossilised birth-death process

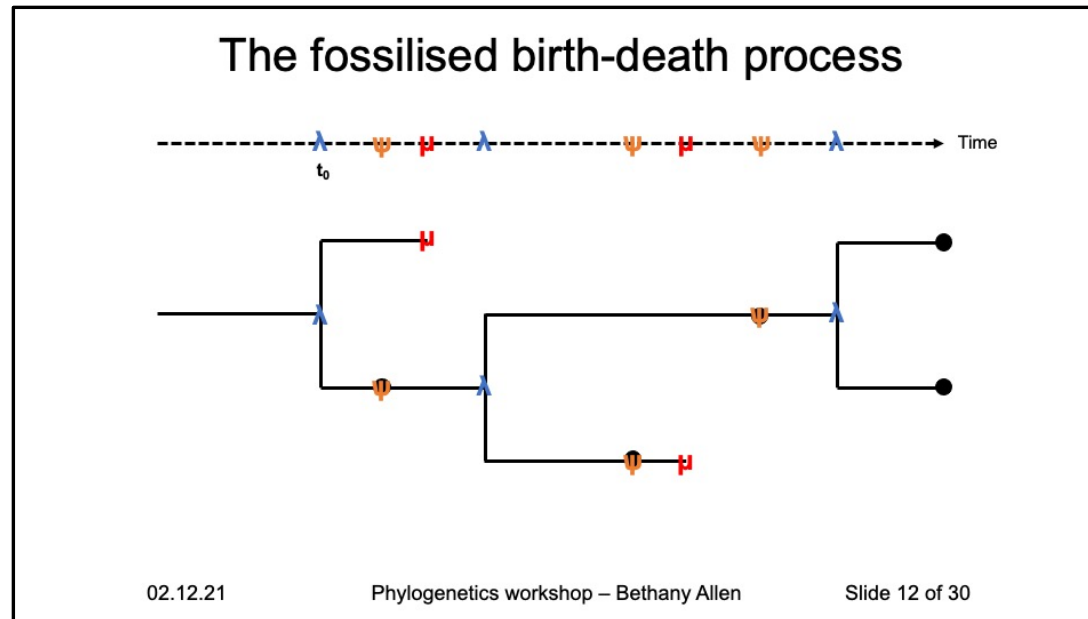


02.12.21

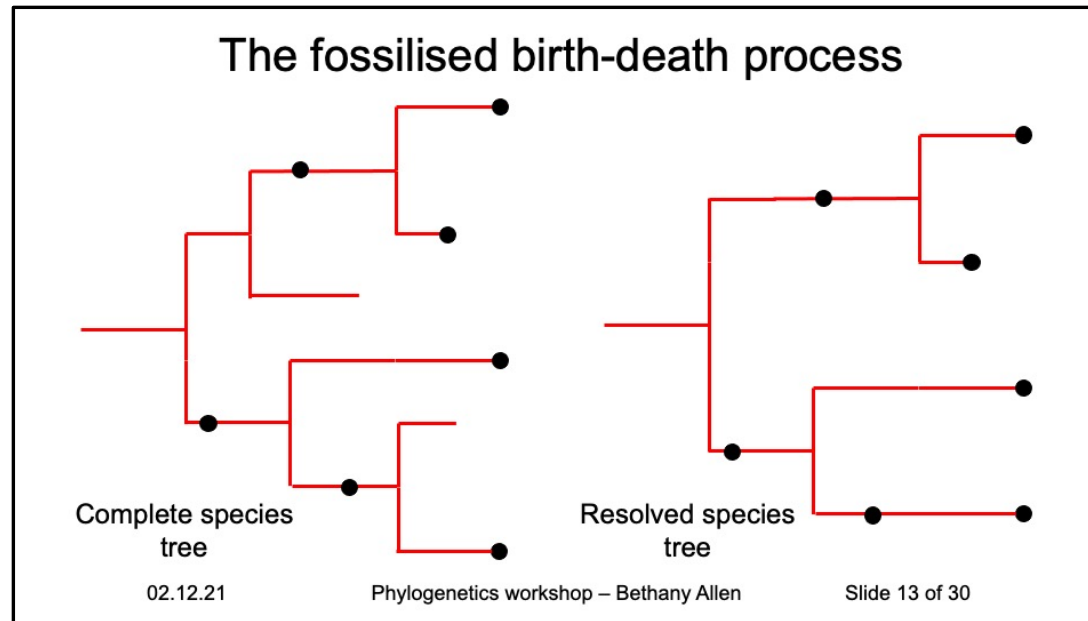
Phylogenetics workshop – Bethany Allen

Slide 11 of 30

For each iteration in the analysis, a value is drawn from each of the parameter distributions.



These parameter values are then used to simulate a phylogeny. If we imagine the flow of time on from the root, speciation, extinction and fossilisation processes occur at their given frequencies (as sampled from the priors), and are allocated to a random branch in the phylogeny. This continues until we reach the present, when we sample our remaining tips with our extant sampling probability. Hopefully this illustrates how simply implementing these rates can create a phylogeny against which we can test the fit of the data.



The simulated tree is created from the evolutionary processes, and so we call this the “complete” species tree. This does not reflect what we might be able to observe, so we then trim away all of the lineages with no fossil or extant samples, to see what the tree would look like following sampling biases; this is the “resolved” tree.

The fossilised birth-death process

Resolved species tree is evaluated against the data (with substitution and clock models for branch lengths)

02.12.21 Phylogenetics workshop – Bethany Allen Slide 14 of 30

This resolved tree is the one we can now compare against our sequence and fossil age data, meaning that incomplete sampling is taken into account within the model. This comparison is done via a substitution model and clock model for branch lengths, both of which I mentioned at the end of Part One.

The fossilised birth-death process

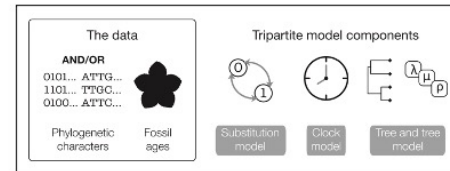
posterior

$$P(\mathcal{E}_{\mathcal{C}}^{\mathcal{A}_{\mathcal{U}} \mathcal{P}} \mid \text{data}) =$$

probability of the character data given everything else* probability of the timetree given the timetree model priors on model parameters

$$\frac{P(\text{data} \mid \mathcal{E}_{\mathcal{C}}^{\mathcal{A}_{\mathcal{U}} \mathcal{P}}) P(\mathcal{E}_{\mathcal{C}}^{\mathcal{A}_{\mathcal{U}} \mathcal{P}} \mid \text{model}) P(\mathcal{A}_{\mathcal{U}}) P(\mathcal{P}) P(\text{model})}{P(\text{data})}$$

marginal probability of the data



Warnock & Wright (2020)

02.12.21

Phylogenetics workshop – Bethany Allen

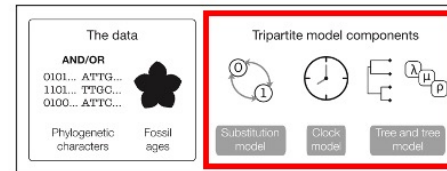
Slide 15 of 30

This is the equation I showed you earlier, modified to fit this scenario. The prior includes all of the model parameters, and also the probability of the phylogeny given the model parameters.

The fossilised birth-death process

posterior

$$P(\mathcal{E}_C^{\mathcal{A}_U \mathcal{P}} \mathcal{O} \mathcal{C} \mid \begin{smallmatrix} 0101\dots \\ 1101\dots \\ 0100\dots \end{smallmatrix} \star) =$$



probability of the character data given everything else*

probability of the timetree given the timetree model

priors on model parameters

$$P(\begin{smallmatrix} 0101\dots \\ 1101\dots \\ 0100\dots \end{smallmatrix} \mid \mathcal{E}_C^{\mathcal{A}_U \mathcal{P}} \mathcal{O} \mathcal{C}) P(\mathcal{E}_C \mid \star \mathcal{A}_U \mathcal{P}) P(\mathcal{A}_U) P(\mathcal{O}) P(\mathcal{C})$$

$$P(\begin{smallmatrix} 0101\dots \\ 1101\dots \\ 0100\dots \end{smallmatrix} \star)$$

marginal probability of the data

Warnock & Wright (2020)

02.12.21

Phylogenetics workshop – Bethany Allen

Slide 16 of 30

I just want to quickly highlight that this equation includes the tree model (here the fossilised birth-death process), a substitution model and a clock model. Methods that include these three components are commonly known as “tripartite” models.

The fossilised birth-death process

Markov chain Monte Carlo (MCMC) is used to explore parameter (and tree) space

This involves sampling from the priors, evaluating the fit of the outcomes to the data, and subsequently exploring nearby parameter space

Most Bayesian analyses use MCMC. Monte Carlo refers to the process of sensitivity testing by selecting values repeatedly from prior distributions, while the Markov chain refers to the fact that each step is based on the previous, but otherwise the process is memoryless (it can keep revisiting the same regions of parameter- and treespace). Sometimes the process selects parameter sets which are a worse fit than the previous; this is to enable the process to escape certain pockets of parameter space in case another, better region exists further away. Various aspects of the process can be fine-tuned to alter the exploration of parameter space if necessary.

The fossilised birth-death process

Subsequent extensions include

- reconstruction of speciation and extinction rates, and lineages (diversity) through time
- varying parameters through time
- dealing with sampled ancestors
- using fossils with uncertain placement in the tree to influence node ages

Various extensions have since been added to this model which give it more flexibility and make it more likely to reflect real evolutionary processes. The first point will be covered in Part Three.

The fossilised birth-death process



Beast2

Bayesian evolutionary analysis by sampling trees

Bouckaert et al. (2019)

Find tutorials, package explanations and
more at **taming-the-beast.org**

02.12.21

Phylogenetics workshop – Bethany Allen

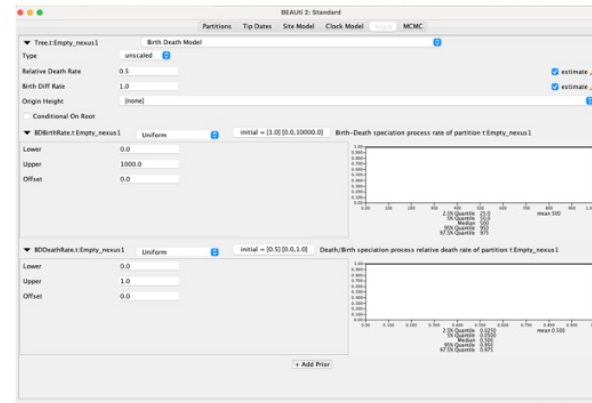
Slide 19 of 30

The best implementation for running the fossilised birth-death process is probably in BEAST. BEAST is a program which can run many different Bayesian phylogenetic analyses. BEAST analyses can easily be transferred to supercomputer clusters, and are supported by tutorials, including those developed by Taming the BEAST.

The fossilised birth-death process

BEAUti is a GUI (graphical user interface) for creating .xml files to run in BEAST

Data, models and priors are entered here



02.12.21

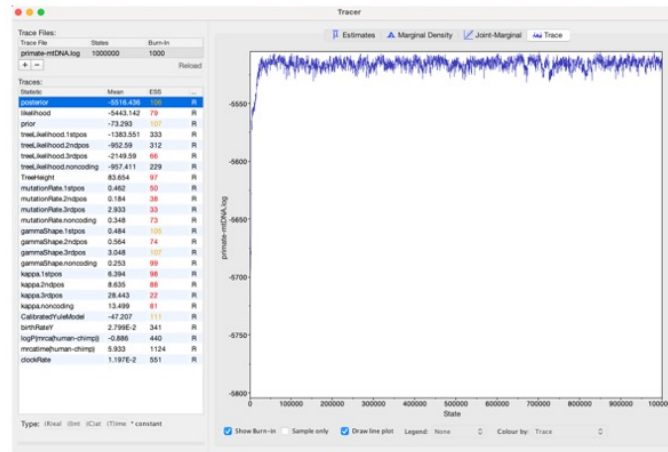
Phylogenetics workshop – Bethany Allen

Slide 20 of 30

Another nice thing about using BEAST is that there is a GUI available in which you can easily design your analysis. BEAUti formats all of your data and models for you, and also visualises all of your parameter distributions, which can really help you understand exactly how they work.

The fossilised birth-death process

Outputs (for any Bayesian analysis) can be viewed and summarised in Tracer, including flexible burn-in



02.12.21

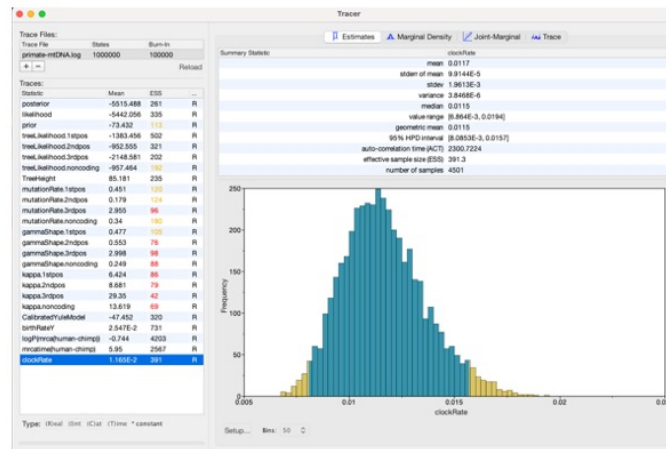
Phylogenetics workshop – Bethany Allen

Slide 21 of 30

Once the analysis is done, it needs to be processed and considered – what does it all mean? Tracer is a good program for this. The early part is discarded; you can see from the trace here that this corresponds to time when the chain is searching through space which is not yet a good fit for the data, which we don't want to include in our final results. This is called burn-in, and can be estimated by eye (trim everything before the posterior probability reaches a plateau), but is often arbitrarily set to the first 10% of the analysis.

The fossilised birth-death process

Outputs (for any Bayesian analysis) can be viewed and summarised in Tracer, including flexible burn-in

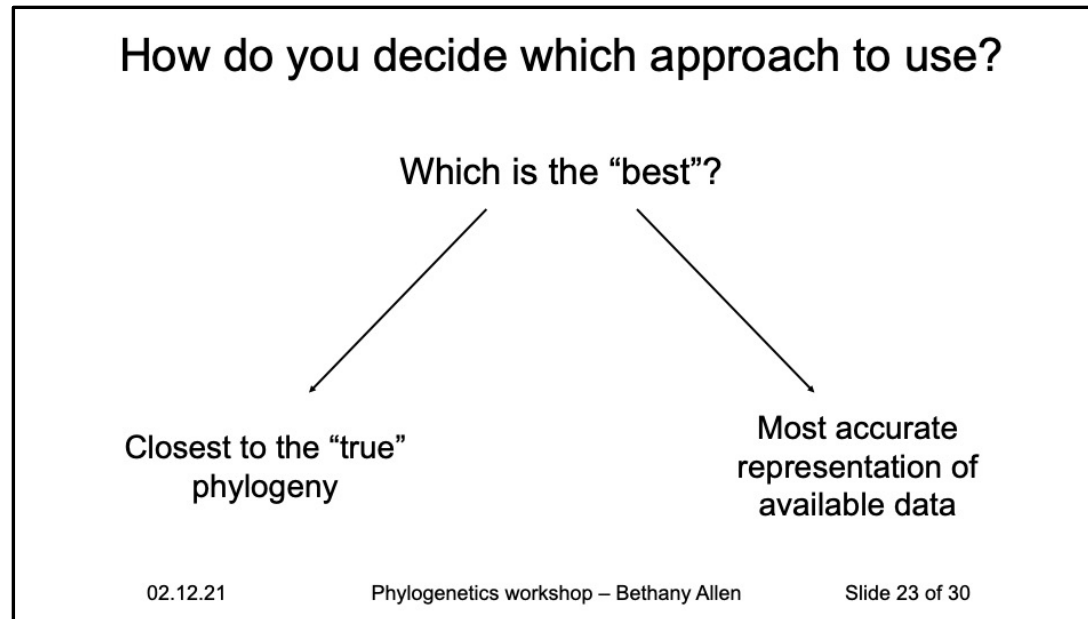


02.12.21

Phylogenetics workshop – Bethany Allen

Slide 22 of 30

We can then look at the parameters associated with our “good” trees. This example shows you the posterior distribution for the clock rate in a fixed clock analysis. Having this distribution provides a mean value but also gives uncertainty around this.



Having learned a little about the main three approaches to evaluating phylogenies, you might be hoping that I can now tell you which is the best. However, I can't. Attempts to determine which is the most accurate or precise have rarely provided much insight. In part, this is because even the goal is uncertain. Do we consider that the “best” approach is the one that gives us a phylogeny closest to the true evolutionary history? We don't know this for any real clade, and can only hope to address this via simulations. It is also fair to consider that the “best” approach is the one that gives us phylogenies that best reflect the data and models they are provided with. But even this is very difficult to determine.

How do you decide which approach to use?

What data do you have access to?

Which philosophy best fits your research question?

How much computational power do you have?

How important is understanding uncertainty?

Instead, it is reasonable to choose which analysis to carry out based on your research question and the data available to you. What implementations exist that fit your data types? Models are applied in each approach, and it is important that those models adequately fit your data types and what you are trying to achieve. The approaches require very different amounts of computational power, and it is valid to make choices based on what you can access. The importance of uncertainty, and the elements of your analysis in which it matters the most, are unique to the research question, and it is sensible to consider whether a particular approach will provide you with this information.

How do you decide which approach to use?

Implementations

Parsimony

Mesquite, PAUP, PHYLIP, TNT

Maximum Likelihood

IQ-TREE, PAML, PhyML, RAxML, Treefinder

Bayesian

BEAST, MrBayes, RevBayes

https://en.wikipedia.org/wiki/List_of_phylogenetics_software

An introduction to some of the more popular implementations of each method – these are a good place to start if you want to research building your own phylogeny. Wikipedia has a fairly complete list of software, which can be really useful if you want to learn more.

How do you decide which approach to use?

	Parsimony	Maximum likelihood	Bayesian
Easy to implement	Yes	Mostly	No
Amount of data incorporated	Small	Medium	Large
Computational power needed	Small	Medium	Large
Size of phylogenies possible	Large	Medium	Small

02.12.21

Phylogenetics workshop – Bethany Allen

Slide 26 of 30

To finish the session, here is a summary of the different approaches and their positives and negatives. Approaches that use more complex methods tend to require a better knowledge of the literature and are more tricky to physically implement. However, they also include more of the available data in their analyses, and might be argued to be superior as a result. This can also be a negative though, as more time and care is needed in compiling the necessary datasets. Bayesian approaches typically need to be run on a supercomputer cluster, whereas the other two can usually be run on a standard computer. The added complexity also means that Bayesian approaches can only handle fairly small phylogenies, which can pose a problem if your question requires lots of tips.

How do you decide which approach to use?

	Parsimony	Maximum likelihood	Bayesian
Major assumption	Evolution is slow	Model choices are correct	Model choices are correct
Can calculate branch lengths within same analysis	No	No	Yes
Provides information on uncertainty	No	Sort of	Yes

02.12.21

Phylogenetics workshop – Bethany Allen

Slide 27 of 30

The philosophy of the approaches should be an important part of your considerations. Assuming that evolution is slow, as in parsimony, might better fit morphological data, whereas the other two approaches require substitution models which are easier to implement for molecular data. The importance of branch lengths to your research question is also important, and it's worth considering whether determining them within the same analysis would be beneficial. Finally, Bayesian analyses provide a good understanding of uncertainty around phylogeny shape and parameter values, whereas the other approaches don't really do this.

Find out more

GitHub: <https://github.com/bethany-j-allen>

Introduction to phylogeny manipulation in R

List of useful phylogenetics programs

Find out more

The fossilised birth-death process, Stadler (2010):

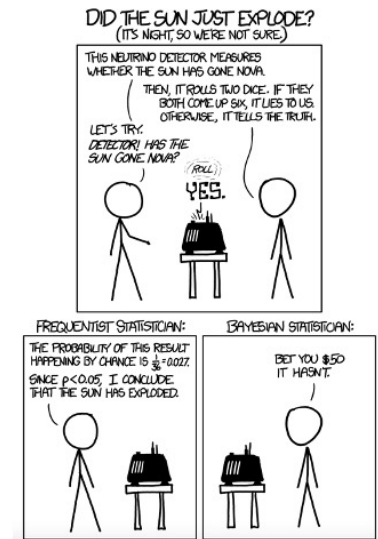
doi.org/10.1016/j.jtbi.2010.09.010

Taming the BEAST, Barido-Sottani et al. (2018):

doi.org/10.1093/sysbio/syx060

Tripartite Bayesian approaches, Warnock & Wright (2020):

doi.org/10.1017/9781108954365



Frequentists vs. Bayesians
xkcd