

**Summary
and
Recommendations**

Avoid the “Black Box”

- Researchers invest considerable resources in producing molecular sequence data**
- They should also be ready to invest the time and effort needed to get the most out of their data**
- Modern phylogenetic software makes it easy to produce trees from aligned sequence data but phylogenetic inference should not be treated as a “black box”**

Choices are Unavoidable

- **There are many different phylogenetic methods**
- **Thus the investigator is confronted with unavoidable choices**
- **Not all methods are equally good for all data**
- **Although we need not understand all the details of the various phylogenetic methods, an understanding of the basic properties is essential for informed choice of method and interpretation of results**

Data are not Perfect

- **Most data includes misleading evidence of relationships and we need to have a cautious attitude to the quality of data and trees**
- **Data can be subject to both systematic biases and noise that affect our chances of getting the correct tree**
- **For example:**
 - **Saturation (noise)**
 - **Alignment artefacts**
 - **Base compositional biases (e.g. thermophilic convergence)**
 - **Branch length or rate asymmetries leading to long branch attractions**
- **Different methods may be more or less sensitive to some of these problems**

Alignment - Homology

- **The data determines the results**
- **The alignment determines the data (hypotheses of homology)**
- **Be aware of potential alignment artefacts**
- **If using multiple alignment software, explore the sensitivity of the alignment to variations in the parameters used**
- **Eliminate regions that cannot be aligned with confidence**

Models

- **Simple models (in ML and distance analyses) often perform poorly because the data does not fit the model**
- **Explore the data for potential biases and deviations from the assumptions of the model**
- **Be prepared to use more complex models that better approximate the evolution of the sequences and therefore might be expected to give more accurate results**

Choice of Models

- **More complex models require the estimation of more parameters each of which is subject to some error**
- **Thus there is a trade-off between more realistic and complex models and their power to discriminate between alternative hypotheses**
- **By comparing likelihoods of trees under different models we can determine if a more complex model gives a significantly better fit to the data**

Choice of Method

- **Not all methods deal with all known problems**
- **LogDet is useful when there are strong base compositional biases but does not deal with rate heterogeneity (need to remove invariant sites)**
- **ML with gamma distribution is useful when there are strong rate heterogeneities across sites**
- **Gamma shape and proportions of invariant sites can be estimated from the data**

An Experimental Science

- **Phylogenetics differs from many sciences in its historical focus**
- **The classical experimental method is not applicable**
- **However, we can perform experiments in the analysis of data**
- **Experiments (multiple analyses) help us to understand the behaviour of the data**
- **The only cost is the time invested!**

Some Experiments

- **Vary the included taxa**

You may be able to minimise the effects of biases by appropriate taxon sampling to break long branches or reduce base compositional biases by introducing intermediate taxa

- **Vary the characters included**

You may be able to improve the fit of data to a model by removing the fastest evolving sites or the slowest evolving sites

Is the data any good?

- Explore the data for phylogenetic signal: randomization tests will identify data that cannot be used to generate reliable phylogenetic inferences**
- Be ready to explore data partitions or ways of treating the data - for example in protein coding genes, systematic biases or noise may differentially affect 3rd positions in codons and might be avoided by excluding this data or by translating DNA sequences and analysing amino acid sequences**

Measure support for groups

- Evaluate relationships shown in trees with bootstrap or other resampling techniques**
- Appreciate that such measures may be misleading if the data is misleading (particularly if subject to systematic biases)**
- Explore the sensitivity of these results to methods of analyses - disagreements should limit confidence in results unless they can be explained as a result of undesirable properties of methods/characteristics of the data**

Hypothesis testing

- **Alternative evolutionary hypotheses may be supported by alternative phylogenetic trees**
- **We can test alternative hypotheses by determining if any of the alternative trees are significantly better explanations of the data**
- **Use constrained analyses to find alternative trees**
- **Use SH or other tests to evaluate alternative trees**

Gene trees and species trees

Remember that molecular systematics yields gene trees

Accurate gene trees may not be accurate organismal trees

Gene duplications and paralogy, lateral transfer, and lineage sorting of plastid genomes can produce mismatches between gene and organismal phylogenies

Use congruence between separate gene trees to identify robust organismal phylogenies or mismatches that require further information