Choosing models for phylogenetic analysis

Graham Jones

Talk given 2011-12-16, this version 2012-01-11

1 About me

I did a PhD in pure maths a long time ago. I have mostly worked in the (overlapping) areas of machine vision and pattern recognition, often using methods of statistical inference.

I became interested in biology, especially macroevolution and phylogenetic analysis, a few years ago. I've been working here with Bengt Oxelman since April, developing models for phylogenetic analysis involving genomic allopolyploids.

SLIDE: I was particularly interested in music recognition. So the data I used to work with looked like this, and now it looks like this.

I am not a biologist.

2 Introduction

2.1 Outline

Here is the overall outline of my talk.

Firstly, I will go through the general framework: how we go from data to knowledge.

Then I will go through a long list of assumptions that are used in phylogenetics. I wanted to make the list as complete as I could, because that seems a useful thing for you to be able to look at afterwards.

However, most of them I will deal with very briefly. I will spend most time on the ones that seem most relevant to NGS data. In particular, that includes those relevant for multi-gene data.

2.2 "Philosophy" for this talk

> All models are wrong, but some are useful.
>
> George Box

George Box is a statistician.

I will assume:

Most of you are working with sexually reproducing organisms (so that 'species' makes sense).

The main aim of the phylogenetic analysis is to infer a species tree or network.

There are other reasons to do a phylogenetic analysis, but I can't cover everything.

3 Models and Inference

The model and the method of inference are different things.

3.1 Simple example of statistical inference

Quick reminder about model, prior, likelihood, posterior.

Example. You want to estimate how many seeds in a batch seeds will germinate without planting them all. You plant three, which all germinate. Estimate the proportion of seeds that will germinate.

Model: Assume each seed has the same probability θ of success. Assume the seeds behave independently. (That is, independent Bernoulli trials.)

Data: 3/3

Two methods of inference: MLE and Bayesian with uniform prior.

Likelihood = θ^3.

MLE: $\hat{\theta} = 1$

Bayesian with uniform prior, posterior mean: $\hat{\theta} = 4/5$

$$\frac{\int_0^1 \theta\theta^3 \mathrm{d}\theta}{\int_0^1 \theta^3 \mathrm{d}\theta} = \frac{\left[\theta^5/5\right]_0^1}{\left[\theta^4/4\right]_0^1} = \frac{1/5}{1/4} = 4/5$$

SLIDE:

On the left, priors: neutral, optimistic, pessimistic

In the middle, the likelihood is always the same. Depends on data, model.

On the right, the posterior. Influenced by data, model, prior. Vertical lines show posterior means.

3.2 'Engineering' or 'ad hoc' methods

Can't say much in general because each one is different.

These methods can seem attractive because they seem simple, and their assumptions are implicit. They can also be fast. With lots of data you might be forced to choose between using a quick and dirty method on all the data or a better method on some of it.

Eg neighbour joining.

3.3 Statistical inference

Model is maths (probability theory) and biology. In biology the models are nearly always probabilistic (or 'stochastic') in nature. Are all the seeds the same? (Or can you see just by looking at them that some are more likely to germinate than others?) Do they behave independently? (Probably not if you plant them close together).

Method of inference is maths (statistical theory) and computing.

You must understand the model. Although it is expressed in the language of probability, it is biology. The details matter.

You do not need to understand much about the method of inference. There is no biology here. In an ideal world, where computers are infinitely fast, and programmers write perfect programs, you would need to know very little. In practice you will have to learn some things about this. I won't be talking about them here.

3.4 Bayesian statistical inference

You do need to understand the difference between frequentist and Bayesian methods of inference.

There is biology in the model and the prior.

3.5 Bayesian vs frequentist

(As slide)

HPD = highest probability density set

4 Assumptions

4.1 Models are based on assumptions

In order to construct a model, you must make assumptions. The assumptions are biological in nature.

Remember the quote: all models are wrong, but some are useful. You will have to make assumptions which you know are wrong, in order to get anywhere.

The question you should ask is NOT: is the assumption wrong, but rather: is the assumption so badly violated that the results are messed up?

4.2 Gene trees and species trees

The assumptions that seem most relevant to this course are ones relating to multiple gene trees. As people say, "every gene tells a different story."

This is one of the things that people are finding as more data becomes available. The world turns out to be more complicated than was previously assumed.

4.3 The list

Next I will go through the assumptions underlying the various models that have been proposed for the evolution of molecular sequences.

I have divided the assumptions into four groups.The division into groups is a bit arbitrary.

I think the second group will be the most interesting over the next few years. The first ones are either safe assumptions, or seem too hard to manage without. The third group are already dealt with in several programs, and you are probably already familiar with them.

5 Nearly always assumed

5.1 Common descent

Do all the organisms have a common ancestor?

When I first started writing this talk, I was thinking this was too obvious. After all, I didn't think there would be any creationists in the audience!

But maybe you have a diverse bunch of viruses.

5.2 Mutations along different branches are independent

Could co-evolution of parasites and hosts violate this?

I don't know! I'm not a biologist remember.

5.3 Constant substitution model

This is the assumption that ratios of different substitution rates (C→T)/(G→T) don't change.

In deep phylogenies, they do. (I have seen this even among primates.)

5.4 Sites evolve independently

Not true, but what to do?

We don't know how to relax this assumption without taking huge amounts of time.

5.5 Genes are units

Has there been merging or fragmenting of genes, or exon shuffling?

Presumably this is a problem in deep phylogenies,

6 Very common assumptions

These are very common, but it is also quite common for one or two of them NOT to be assumed.

These seem the most interesting group. A lot of work is going on with these, but there is even more to do.

6.1 The alignment is correct

This assumption is an oddity. It is not just an assumption about the evolutionary process, but one about the evolutionary process PLUS a previous computation. An alignment is not an observation. Statistical theory doen't cover the use of things like this.

This assumption is a bit of engineering or ad hoc processing which has been inserted into a mainly statistical method.

If there are lots of insertions and deletions, the assumption can be seriously violated. There is some work being done on this, such as BALi-phy.

6.1.1 Relevant software

BALi-phy http://www.biomath.ucla.edu/msuchard/bali-phy/

6.2 Genes are orthologous

Is there any gene duplication or loss?

Is there horizontal gene transfer?

6.2.1 Relevant software

PrimeGSR http://prime.sbc.su.se/primeGSR/

[[Look up DLRS, DLTRS, Bengt Senn?]]

PHYLDOG (fast, ignores times I think)

ARG in BEAST (Marc Suchard)

6.3 Speciation is tree-like (binary splits)

With LGT at least the species tree is a tree. With hybridization, the species tree becomes a network.

In homoploid hybrids, each gene comes from one or other parental species. In genenomic allopolyploids, the two chromosome sets from the parental species coexist in the hybrid.

6.3.1 Relevant software

Meng, Kubatko, Theoretical Population Biology, 2009

Mine, one day! (for allopolyploids).

6.4 Species are clearly delimited

Can each organism be unambiguously assigned to a species? (What is your definition of species?)

6.4.1 Relevant software

BPP Yang Z., Rannala B. 2010. Bayesian species delimitation using multilocus sequence data. Proc. Natl Acad. Sci. USA 107, 92649269. (doi:10.1073/pnas.0913022107).

6.5 Speciation is instantaneous

Related to above (6.4). How long does it take for one species to become two?

There may be a period when gene flow between two subpopulations is limited but still occuring at a low rate.

6.5.1 Relevant software

Isolation with Migration IM http://genfaculty.rutgers.edu/hey/software

Hey J. 2010. Isolation with Migration Models for More Than Two Populations. Mol Biol Evol 27: 905-20

6.6 Lineage sorting is complete

Is the effective population as big as the length of the branches measured in generations?

On the left there is a deep coalescence, but the gene tree still has the same topology as the species tree. On the right, the topology is different.

If you have enough genes you can hope that most of them agree with the species tree. More individuals from each species also helps.

Don't forget that the spcies tree is usually unknown too, and is often the main thing you wnat ot estimate.

Also note that this problem can occur at the same time as the others I've just mentioned are violated:

- Genes are orthologous
- The evolutionary process is tree-like
- Species are clearly delimited
- Speciation is instantaneous

6.6.1 Relevant software

*BEAST Heled and Drummond, Mol Biol Evol, 2010

BEST

6.7 Genes are recombinational units (for coalescent models)

Here we are looking inside a single gene tree.

An assumption for coalescent and multi-species coalescent models.

6.8 Substitution process is time-reversible

Nobody believes this assumption. Does it matter if it is violated? Is it useful for rooting trees?

(I think Johan Nylander has some relevant references for this.)

6.8.1 Relevant software

ComplexSubst model in BEAST.

7 Variable assumptions

These are typically available as options in programs like MrBayes and BEAST.

7.1 Strict/relaxed clocks, nonclocklike

In MrBayes, can use clocklike or non-clocklike. Clocklike goes with rooted topologies. Non-clocklike goes with unrooted topologies.

In BEAST, all topologies are rooted. Clocks can be strict or relaxed.

7.2 Partitions, linked/unlinked parameters

What should the partitions be? Genes? Divided into codon units?

What parameters should be shared (linked) between partitions? (tree topology and times? site rate? substitution model?)

This is the sort of problem that biologists ask me, but there isn't a mathematical answer to it. There might be an answer (of sorts) if you understand the molecular biology (which genes/regions likely to be similar, etc).

7.3 Site rate heterogeneity

Which model? Gamma distribution?

RaxML has a category model which is interesting

7.4 Relative substitution rates

Which model? HKY and GTR most popular

Usually need to assume something about the root, eg nucleotide frequencies

8 Priors

In a Bayesian approach, you also need to make assumptions about the prior distributions of all the parameters.

8.1 Tree (or network) prior for topology and node times

Birth-death models, fossil calibration, fixed clades.

8.2 Prior for relative rates of partitions

Lots of data tends to mean lots of partitions.

You need to assume something about the relative clock rates.

People usually use the very diffuse 1/X prior. I guess a less diffuse prior would be more realistic and might give more accurate results. (But might make little difference.)

8.3 Branch rates / branch lengths prior

The most popular are exponential branch lengths in MrBayes, uncorrelated lognormal branch rates in BEAST.

There are several papars indicating this is important to get right. (eg "Lost in the land of long trees", Syst. Biol.)

8.4 Relative substitution rates prior, Site rate heterogeneity prior, Root frequencies prior

Eg: prior for kappa in HKY model.

Eg: prior for shape parameter for gamma distribution

Eg: Dirichlet prior for frequencies.

9 Overfitting and underfitting

Overfitting means the model is too flexible, so that the model fits the noise rather than the signal in the data. Underfitting means the model is to rigid, and cannot fit the signal.

9.1 MLE (frequentist)

Maximum likelihood estimators and AIC/BIC.

9.2 Bayesian

Here, lots of parameters can be used - IF you can express a sensible prior.

10 Conclusion

Statistical inference is based on models. Models are based on assumptions. Assumptions are biological.

Use statistical methods if you can, engineering methods if you have to.

Don't just accept defaults in programs. The programmers may be great biologists, but they don't know about your data.