Assessing Model Adequacy in Phylogenetics - Are the tools powerful?

Daisy Shepherd & Steffen Klaere

PhD Candidate, The Department of Statistics, UoA

AASC18, Rotorua

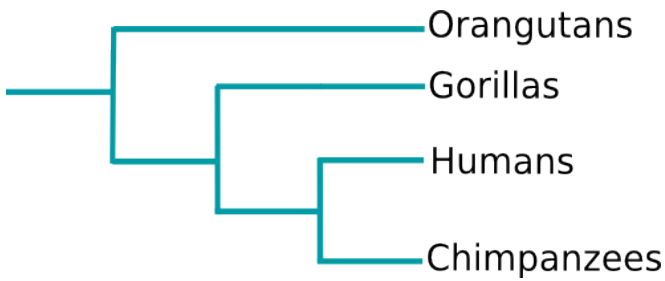
4 December, 2018

SCIENCE
DEPARTMENT OF STATISTICS

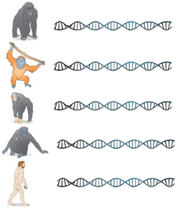
Outline

What is Phylogenetics?

- All organisms have DNA.
- Map the differences in DNA.
- How closely related are these groups?
- **Aim:** Derive their evolutionary history.



Models of Molecular Evolution



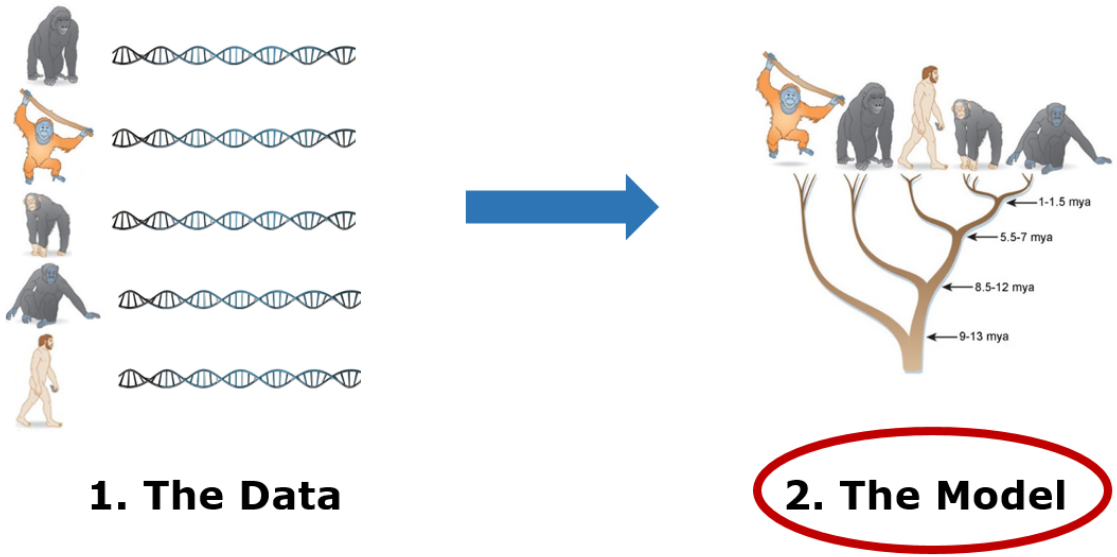
1. The Data

Species 1: AACTTACTACGTACGAT...
Species 2: CACCTATGAGATCGCGA...
Species 3: TAAAAACACTGACACGT...
Species 4: ATAAATATTCCGTGATC...
Species 5: GTGTTTCGATATGCTCG...
⋮
Species k: GTAGGCTACACACATTA...

$$\boldsymbol{X} = \underbrace{\begin{bmatrix} A & A & C & T & T & A & \cdots \\ C & A & C & C & T & A & \cdots \\ T & A & A & A & A & A & \cdots \\ A & T & A & A & A & T & \cdots \\ G & T & G & T & T & T & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}}_{n} \; k$$

Site ⟶ Character "AAATT..."

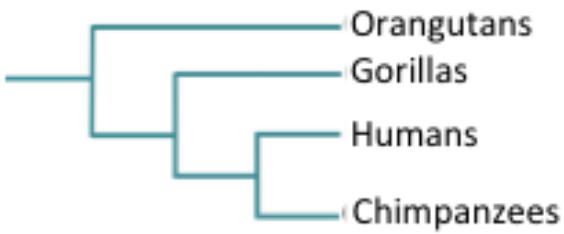
Models of Molecular Evolution



- Model the changes in DNA over evolutionary time.
- Models are **high-dimensional**.
- Parameters explain elements of evolutionary process.
- Lots of methods to select the 'best' model.

Model Adequacy

- **But...**The 'best' fit does not necessarily imply a **good** fit!
- Poorly fitting models lead to poor estimation of evolutionary relationships.



Surely there are tests to check model fit, right?

The Problem

Yes... and No.

Goodness of fit assessment is a critical step...

... but very rarely applied in phylogenetic analysis.

Approaches have been suggested...

... but are hindered by peculiarities of phylogenetic data.

Some approaches seem useful...

... but not implemented in any software.

Current GOF Approaches

- Adapt established GOF statistics to phylogenetic framework.
- Look at deviations between observed (obs) and expected (exp) character counts.

General Statistics:

Deviance: $$G = -2 \sum obs \ (\log(obs) - \log(exp))$$

Pearson X^2 : $$X_P^2 = \sum \frac{(obs - exp)^2}{exp}$$

GOF Assessment

Under good conditions, G and X_P^2 are approximately χ^2_{df} distributed.

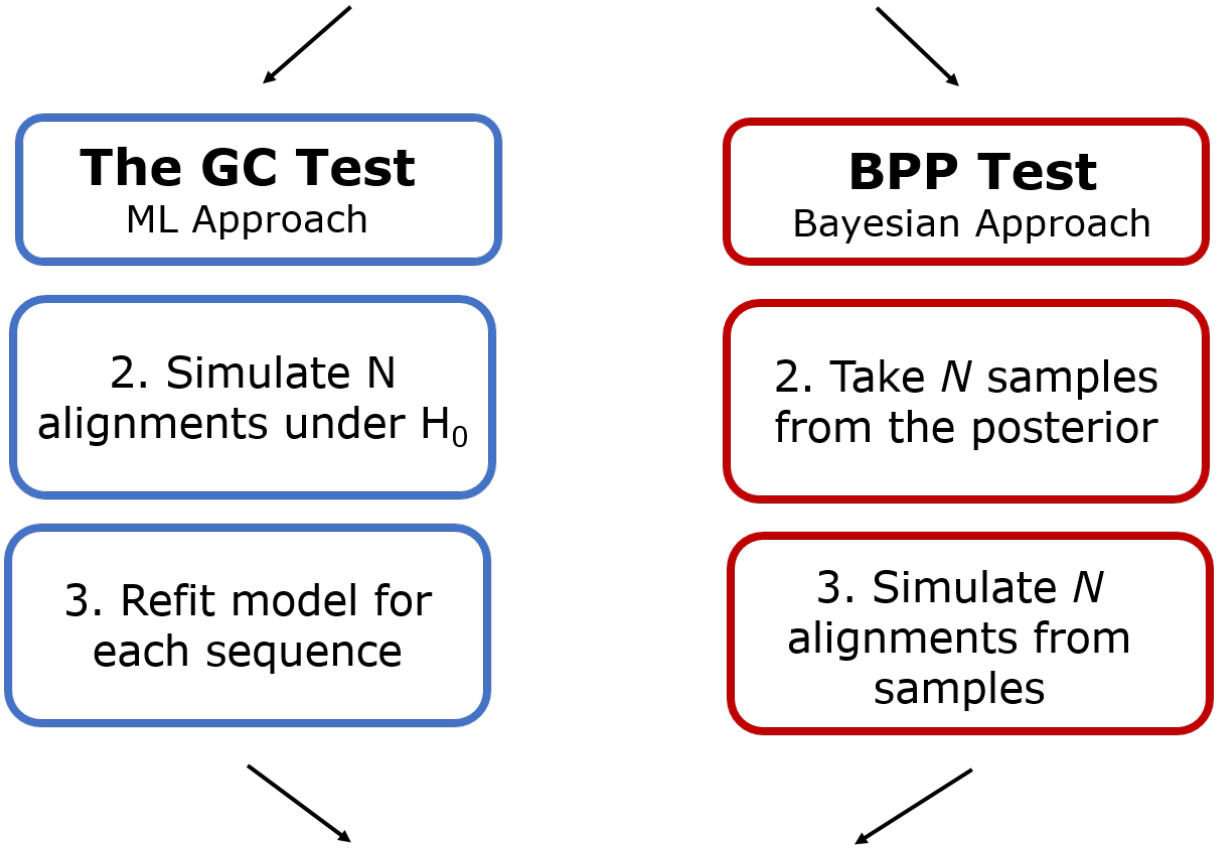
Do these conditions hold for phylogenetic data?

- How do we specify df for the tree T?
- Sparseness in nucleotide data.
- 4^k potential characters - How many do we actually observe?

Conditions do not hold for phylogenetic data!

The solution?

1. Calculate test statistic for **real** alignment



4. Calculate test statistic for each simulated alignment

5. Compare statistic from (1) to distribution of test statistics from (4).

Our study

- ✓ Remove dependency on χ^2 distribution.
- × No implementation in software, rarely used.
- × Do not know a lot about the power of the test.

Aim: Are the tests powerful in the presence of `bad' data?

Data and Methodology

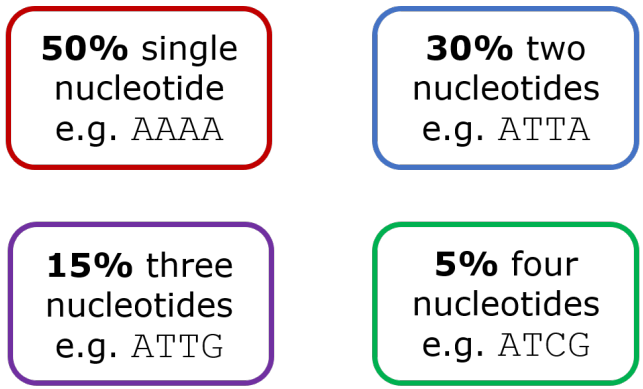
1. The Data:

⇒ Generate under $H_1 \rightarrow$ randomly select nucleotides.

⇒ Simulated 100 alignments for each pairing of k and n.

- Taxa: $k = 10, 20, 30, 50$
- Number of sites: $n = 500, 1000, 5000$

⇒ Keep basic structure of phylogenetic data and sites' nucleotide content.



Data and Methodology

2. Testing:

i. Perform model fitting (ML Analysis in R, Bayesian Analysis in MrBayes).

ii. Calculate test statistic (G and X^2) on the raw data.

iii. Perform the GC test for $N = 100$ replicates.

iv. Perform the BPP test for $N = 100$ replicates.

3. What are we looking for?

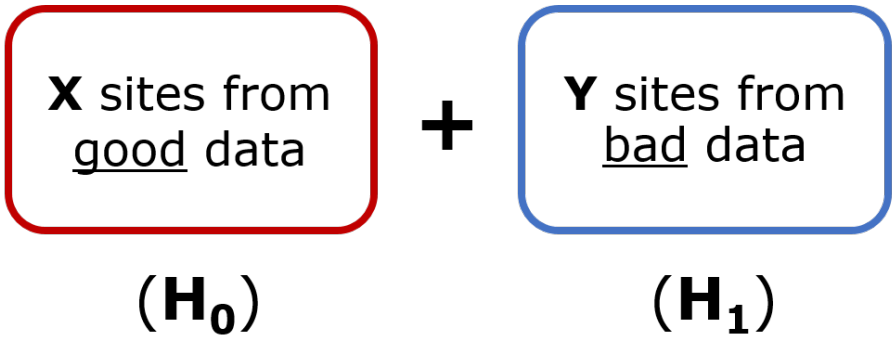
- Testing H_0 : the model fit is adequate.
- Do the tests reject H_0? **Powerful?**
- Are results consistent across test statistic and data dimension?

Results

The GC Test

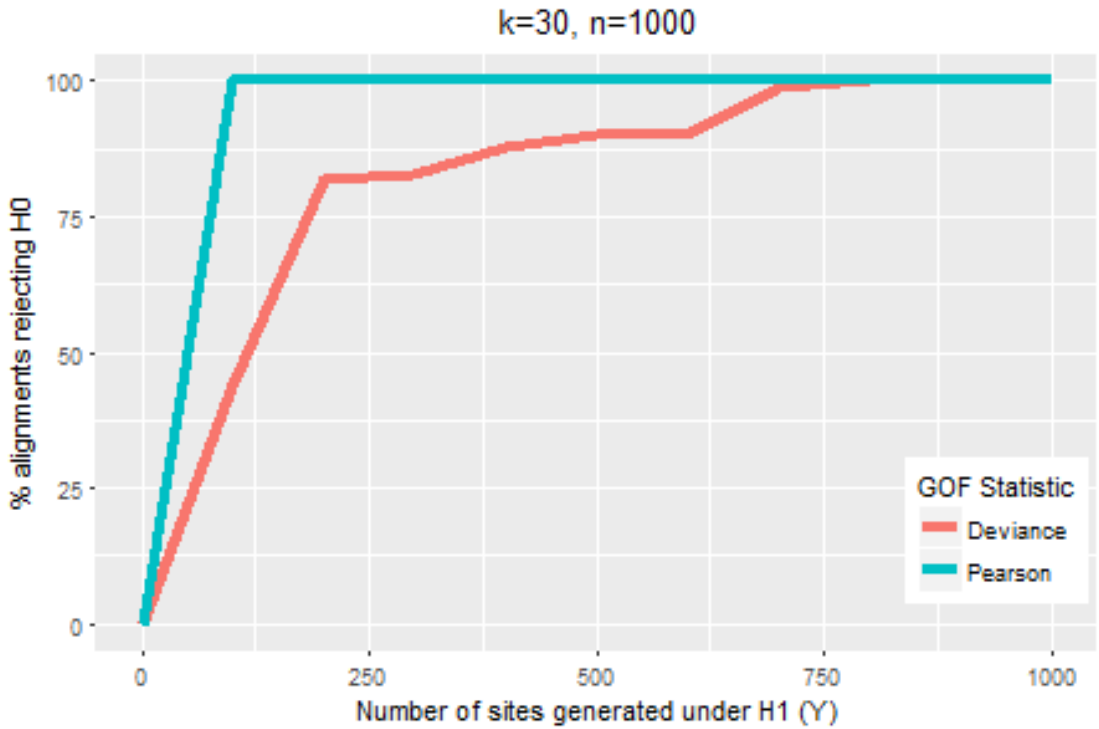
k	n	Deviance		Pearson	
		Reject H_0 P-value ≤ 0.05	Not reject H_0 P-value > 0.05	Reject H_0 P-value ≤ 0.05	Not reject H_0 P-value > 0.05
10	500	1	99	5	95
	1000	31	69	2	98
	5000	99	1	2	98
20	500	89	11	99	1
	1000	99	1	99	1
	5000	100	0	100	0
30	500	100	0	100	0
	1000	100	0	100	0
	5000	100	0	100	0
50	500	100	0	100	0
	1000	100	0	100	0
	5000	100	0	100	0

What degree of ‘bad‘ data for the test to fail?



X	Y
0	1000
100	900
200	800
⋮	⋮
1000	0

What degree of ‘bad‘ data for the test to fail?



- ✓ Pearson is **really** good.
- ✓ Deviance relatively good if $> 25\%$ ‘bad’ data.

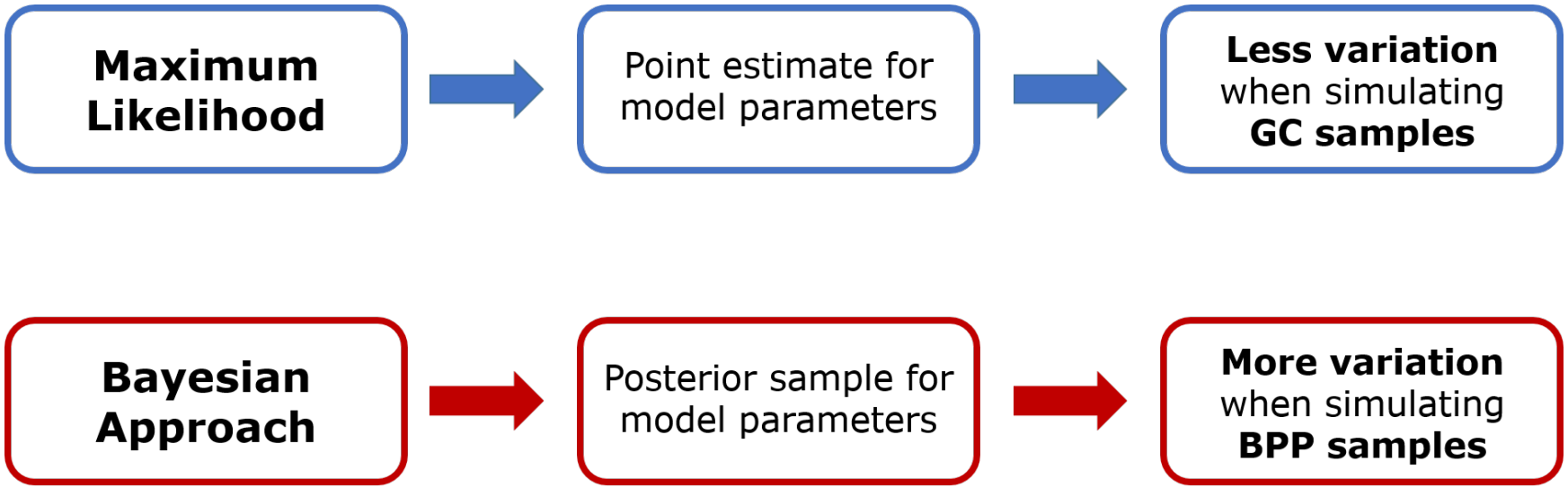
How did the BPP test perform?

Incredibly poorly!

H_0 was consistently **accepted** for both G and X^2 (for all alignment sizes).

All p-values > 0.05.

Why the difference - GC test vs. BPP test?



Conclusions & the Bigger Picture

GC Test:

- ✓ Powerful!
- ✓ Really good with Pearson statistic.
- ✓ Pretty good with deviance.
- ? *Sample size effect.*
- × High computational times.

BPP Test:

- × Not powerful.
- × Garbage in → Garbage out
- ? *Consider the range of posterior.*
- ✓ Quicker computational times.
- ✓ Bayesian analysis very popular.

⇒ **Take home message:** These approaches are promising, so let's start implementing them!