Sliding Through Phylogenetics

Daisy Shepherd

The University of Auckland

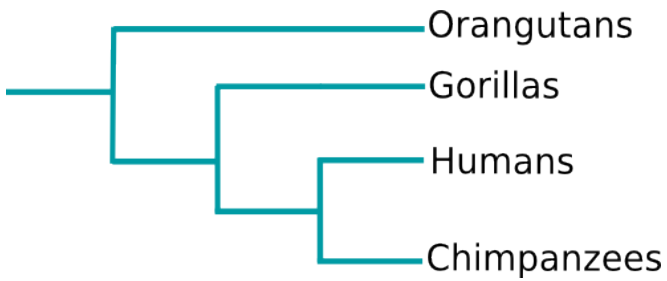
dshe078@aucklanduni.ac.nz

01/12/2015

Outline

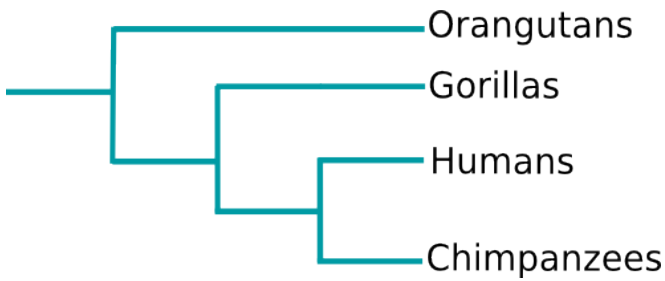
What is Phylogenetics?

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



What is Phylogenetics?

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

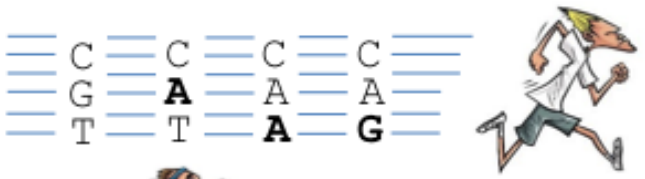


The Problem

- Sites evolve over time in a number of ways.
- Usually assume each site evolves at a similar rate.

However, this is not always the case...

Some sites change *quickly*...



... whilst others change <u>slowly</u>....

... or not at all.

Rates-across-sites

Sites which have the same rate display rate **homo**geneity.

Sites which have differing rates display rate **hetero**geneity.

How do we model these?

Fixed, constant rate

Use γ distribution
γ = rate parameter

So why is this a problem?

- Models need to detect and handle this heterogeneity.
- Poorly fitting models lead to poor estimation of evolutionary relationships.

The Current Approach

- When selecting a model (homogeneous vs. heterogeneous), usually look at the whole alignment.
- Heterogeneous model - one estimate for the rate parameter γ.

Complete alignment analysis

(Traditional approach)
Single analysis

✓ Simple and quick

✓ Common practice, so widely applied in software

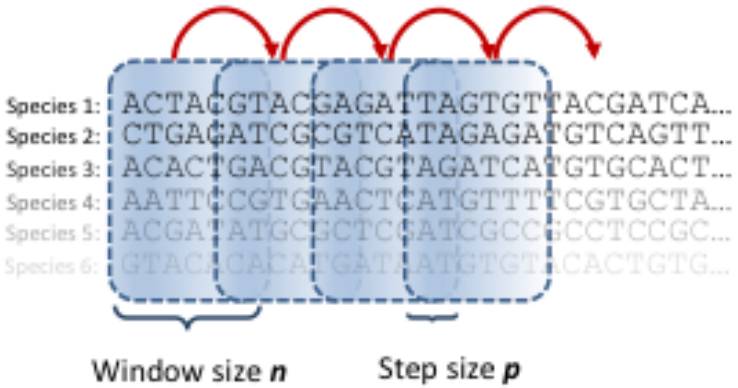
× **Generalises behaviour across all sites**

The Proposed Solution

The Sliding Window Approach

How does it work?

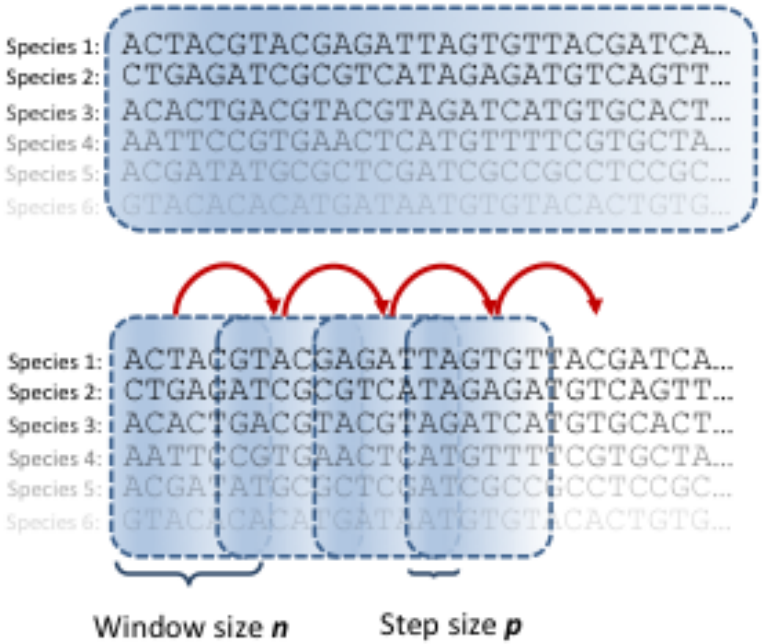
1. Look at the first n sites.
2. Fit a model to only this window of sites.
3. Slide window along p sites, to a new group of n sites.
4. Fit a model to this new window.
5. Repeat until entire alignment has been covered.



Sliding Window analysis

(Our proposed approach)
Multiple analyses

What did we want to do?



Complete alignment analysis

(Traditional approach)
Single analysis

Sliding Window analysis

(Our proposed approach)
Multiple analyses

Aim: Test whether the Sliding Window approach improves our ability to detect variation in evolutionary rates.

Data and Methodology

1. The Data:
Simulated 300 alignments (5000 sites)

- 10, 50 and 100 taxa
- Random γ parameter values used for hetero model
- Generated random topologies
- Random insertion point (for hetero region into alignment)
- Window size $n = 500$ sites
- Step size $p = 50$ sites

Data and Methodology

2. Testing:

i. Perform a complete alignment analysis.

ii. Implement the Sliding Window (SW) approach.

iii. Compare the results.

3. What are we looking for?

- Are heterogeneous or homogeneous rates detected?
- Accurate estimate of the rate?
- Is the SW approach an improvement?

Results

Complete Alignment Analysis

Based on p-values from the likelihood ratio test:

Taxa Number	n	Accept H_0	Reject H_0
10	100	25	75
50	100	20	80
100	100	0	100
H_0: Homogeneous model is the true model			

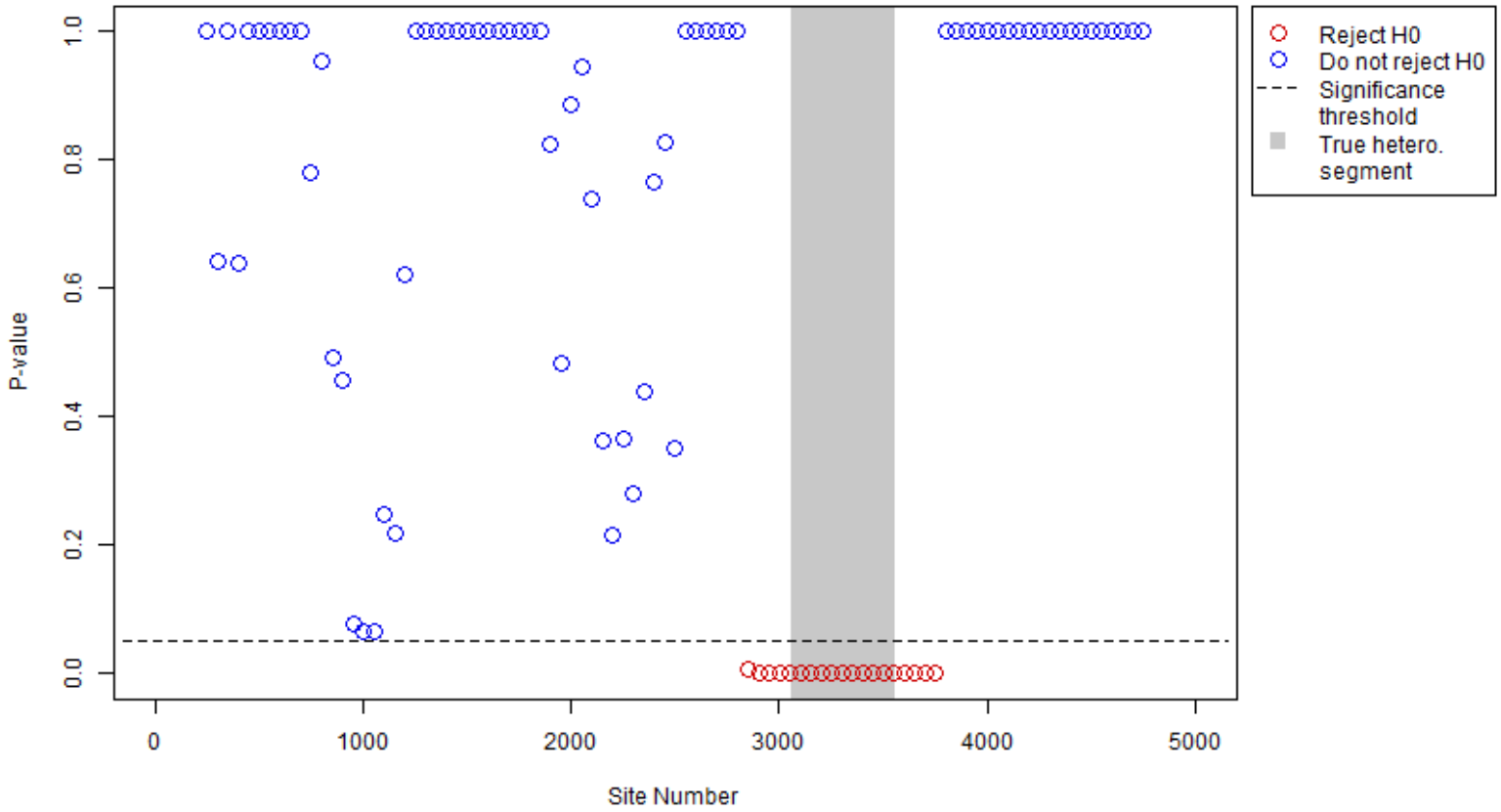
Table : The number of simulations which accepted and rejected the null hypothesis under the complete alignment analysis.

Did not detect varying rates in 15% of the alignments.

1. Were the heterogeneous regions detected?

LRT comparing homogeneous vs. heterogenous model.

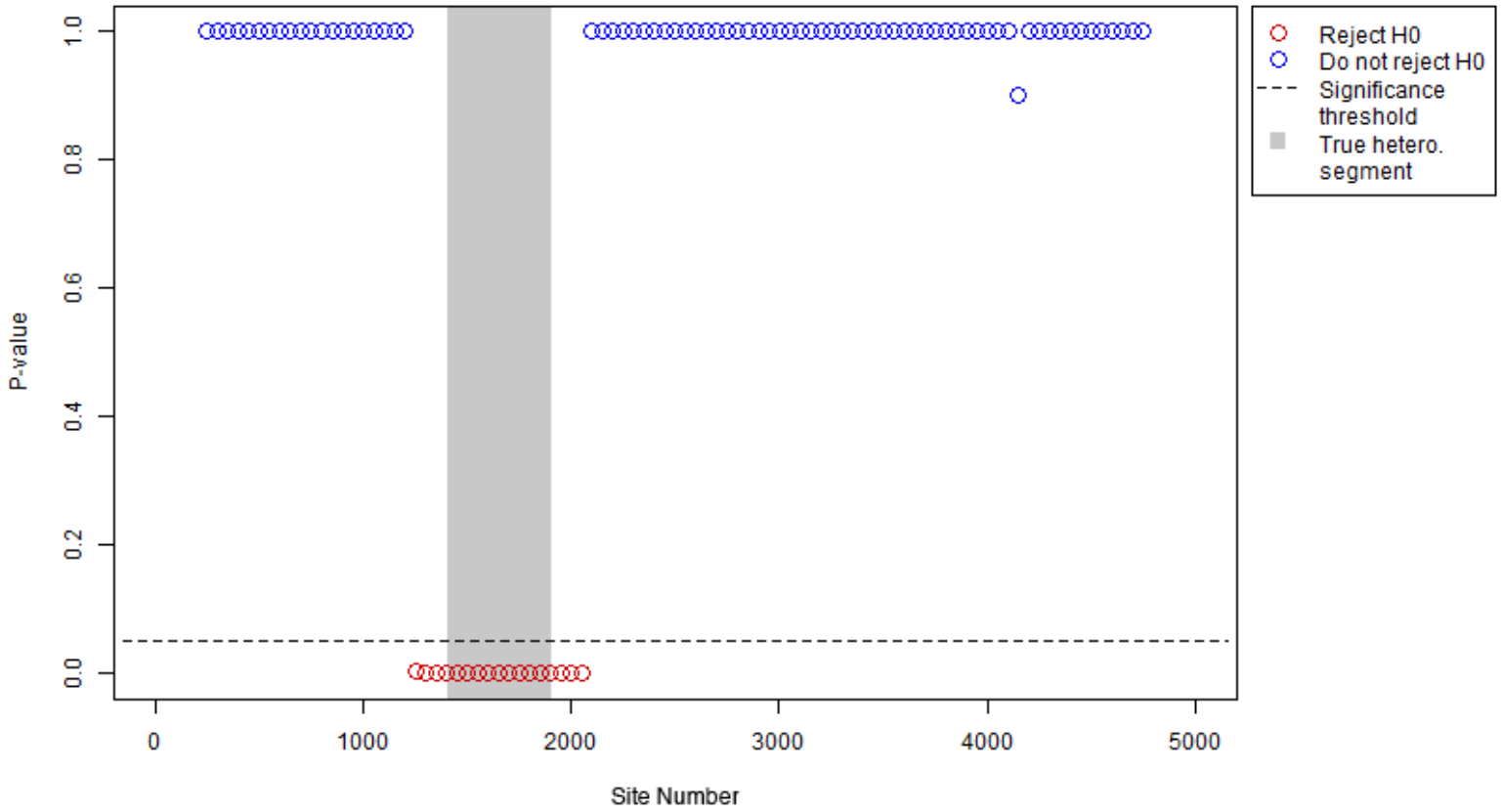
(Complete alignment analysis favoured **hetero**geneous model.)



1. Were the heterogeneous regions detected?

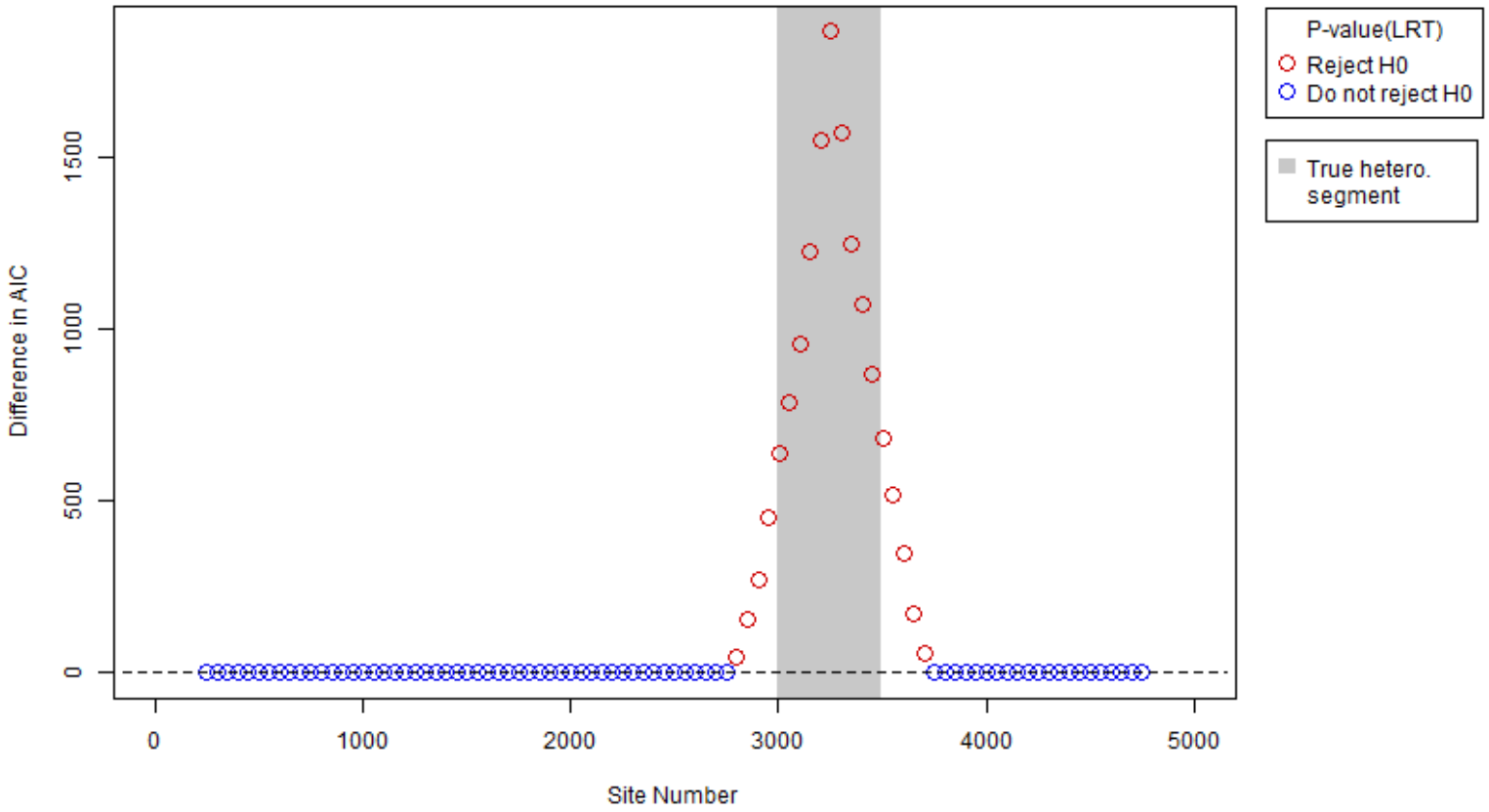
LRT comparing homogeneous vs. heterogenous model.

(Complete alignment analysis favoured **homo**geneous model.)



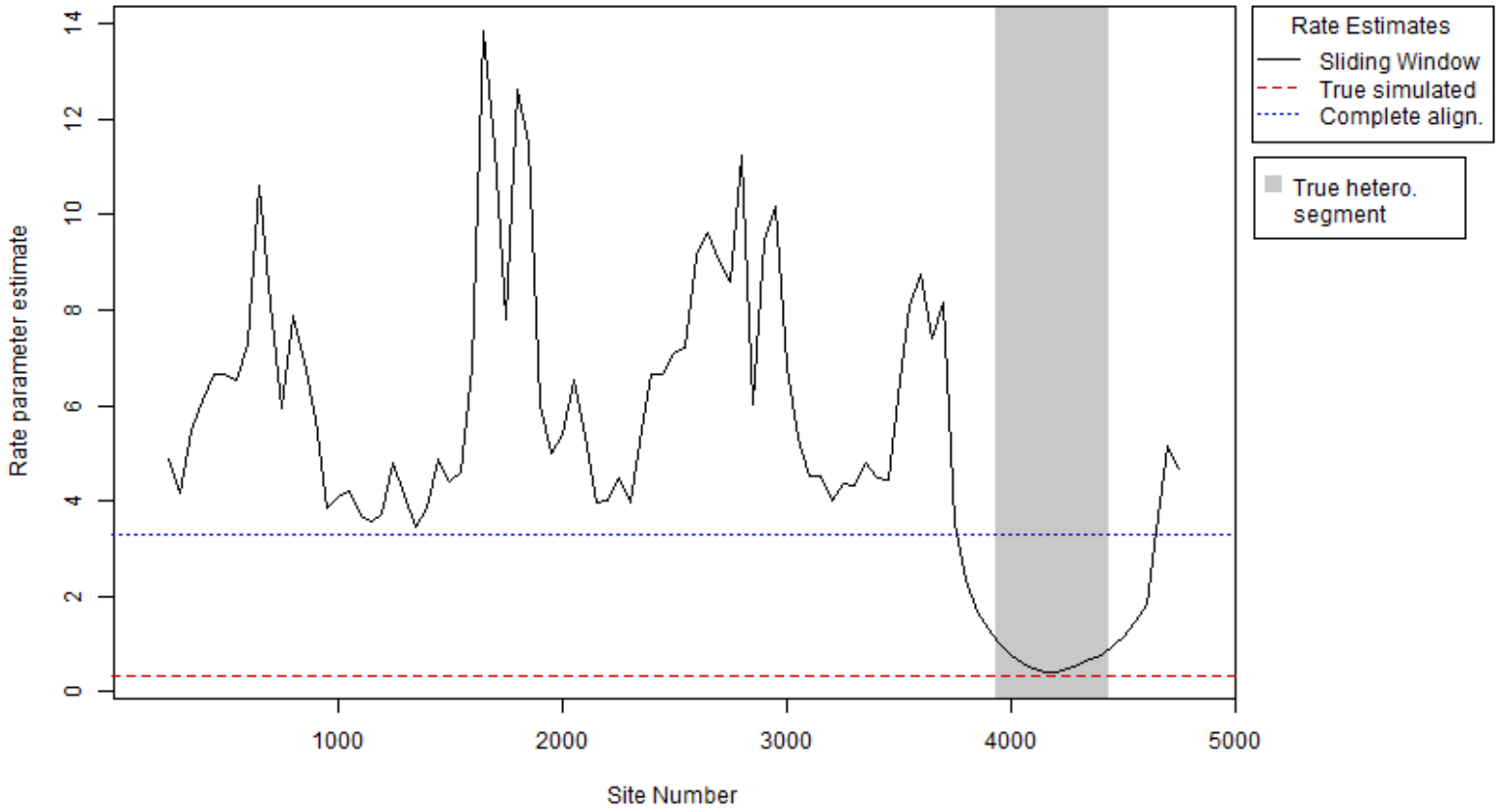
1. Were the heterogeneous regions detected?

Difference in AIC, BIC comparing homogeneous vs. heterogenous model.



2. How accurate was the rate estimate?

Recovered from fitting heterogeneous model.



What did we find?

- ✓ SW approach detected heterogeneous rates consistently.
- ✓ Detected in more situations than under traditional approach.
- ✓ Strong benefits from profiling an alignment.
- ✓ Better overall detection of rate heterogeneity.

Conclusions

Overview: The Sliding Window Approach

Strengths:

- ✓ Gain a deeper insight into the true behaviour of the alignment.
- ✓ Allows multiple analyses which better detect any pattern variation (rates, topology etc.)

Weaknesses:

- × Choosing an appropriate window and step size can be tricky.
- × Computation time can be high.

- SW approach is undoubtedly a useful tool.
- Potential to better detect heterogeneity, and to improve the statistical models we use.

Where to from here?

- $\Rightarrow$ Continue to test the SW approach in other phylogenetic applications (outlier detection, different forms of heterogeneity).
- $\Rightarrow$ Create software to make the SW approach more accessible.
- $\Rightarrow$ Finding optimal window and step sizes?

Acknowledgements

Steffen Klaere

Jessica Leigh