Recombination Breakpoint Detection using the Influence Function

Alethea Rea[1], William Rea[2] and David Bryant[3]

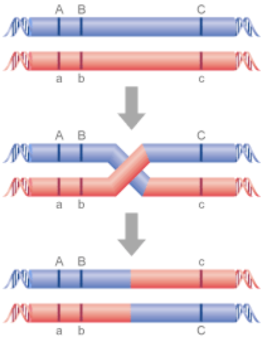
1. Department of Statistics, University of Auckland, NZ
2. Department of Economics and Finance, University of Canterbury, NZ
3. Department of Mathematics and Statistics, University of Otago, NZ

Outline

- Introduction to:
    - Recombination;
    - Influence Function; and
    - ART.
- Simulation study
- Case study: *Candida albicans*
- Discussion

Recombination

- Recombination: A single strand of DNA has at least two origins
- Its presence violates the assumption of a tree



Influence Function

- Quantifies the influence of each site on the tree likelihood calculation.
- The influence function is given by

$$IF(X_h) = \underbrace{l_T}_{\text{Likelihood on whole alignment}} - \underbrace{l_T^{(h)}}_{\text{Likelihood on alignment with a single site removed}}$$

- Sites which have large negative influence function values do not fit the maximum likelihood tree.

(Bar-Hen et. al. 2008)

Atheoretical regression trees

- This technique is from time series and it finds an optimal set of regimes in a time series.
- The algorithm creates nested partitions of segments with the same mean.
- It is based on minimising the sum of squares.
- It is effective for very long series.

Detecting Recombination

- With recombination the site ordering is important.
- The influence function may show recombination as a cluster of large negative values.
- ART would be able to detect a cluster of large negative values as a change in the mean.
- Therefore we apply ART to the influence function.

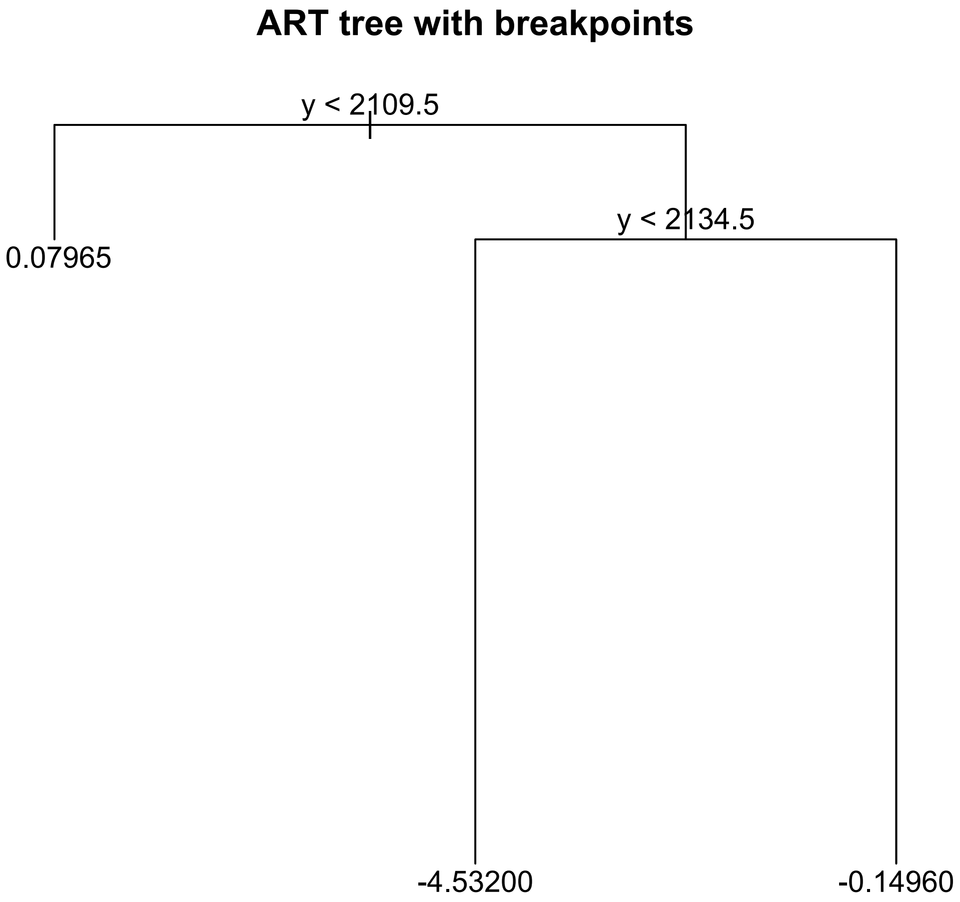
Simulations

- Simulations showed that the influence function with ART was effective in detecting recombination when
    - ✦ sequence divergence was high;
    - ✦ the proportion of the alignment affected by recombination is small; and
    - ✦ the recombination event strongly influences the topology.

Case Study: *Candida albicans*

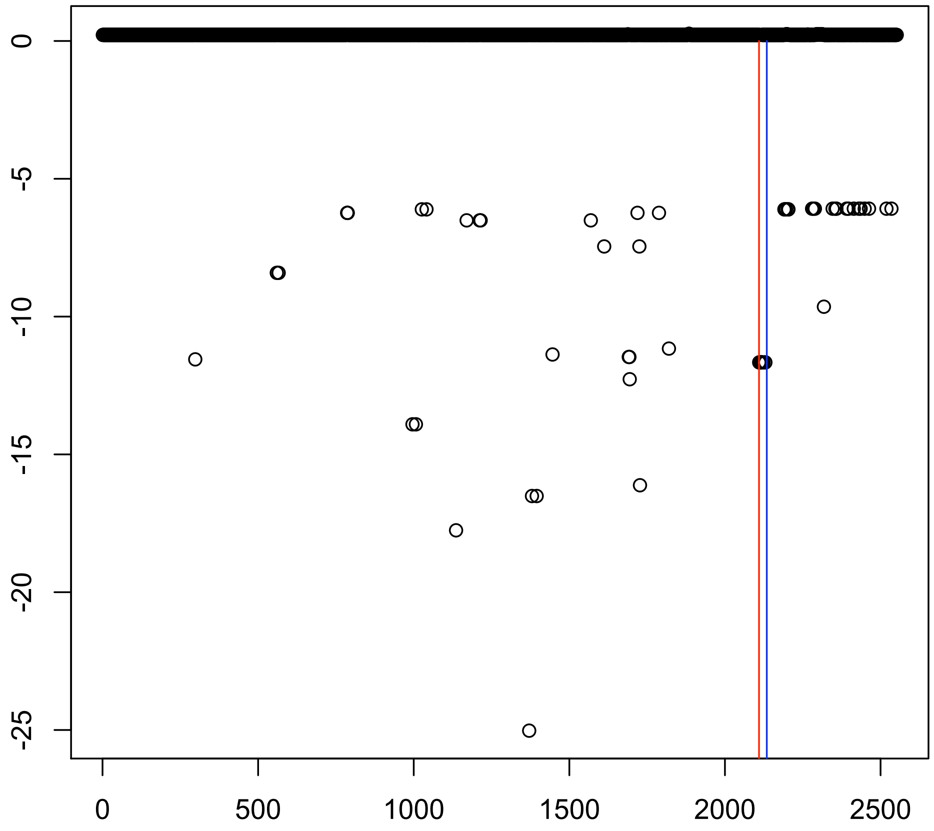
- 45 strains of *Candida albicans* (a yeast).
- 2553 base pairs of mitochondrial DNA.
- The authors investigated clonality and recombination.
- Using maximum parsimony and networks they concluded there was evidence for clonal proliferation and recombination.

(Anderson et. al 2000)



ART tree with breakpoints

Likelhood Influence Function and Breakpoints



Summary

- The influence function may contain information on recombination.
- With ART we can detect the locations of recombination events using the influence function.
- The method is most effective when detecting small recombinant segments within an alignment.

Acknowledgements

- Co-authors: William Rea and David Bryant.
- University of Auckland and the University of Otago.
- TEC for funding my PhD.
- The phylomania organisors.

Tertiary Education Commission
Te Amorangi Mātauranga Matua