Improved Recombination Breakpoint Estimation Utilizing Likelihood Ratio Tests: Like_HMM

Kevin Ziegler[1], and Lemmon A.[1]

[1]Department of Scientific Computing, Florida State University, Tallahassee, United States

Introduction

Recombination causes genomic sites to have differences in their genealogical history. For this reason, recombination can produce error when phylogenetic trees are estimated under models that assume all sites have the same history [1][2][3]. One solution to this problem is to identify and remove loci with a history of recombination. A second solution, more amenable to genome-wide alignments, is to split loci with a history of recombination into two or more separate loci for analysis. A subset of the available methods aim to identify these recombination breakpoints. To our knowledge the accuracy of these methods have not been compared comprehensively. The method presented here automates and extends an existing method for detecting recombination breakpoints in an DNA sequence alignment using a Hidden Markov Model (phyML_Multi, [4]). The performance of our method is compared to existing methods using a simulated dataset designed to emulate Hominidae. Our program uses Likelihood Ratio Tests (LRTs) to substantially reduce the rate at which phyML_Multi falsely predicts recombination break points

Simulation Conditions

To exhibit the functionality of our extension to phyML_multi we simulated a tree mirroring the phylogeny of Hominidae using ms[5]. The divergence times of Hominidae were obtained from timetree.org [6], population sizes were obtained from [7][8], and generation times were obtained from [9]. Scenarios with recombination rates of 0 and 10^{-9} were simulated. Alignments of 10,000 base pairs were simulated using seq-gen[10] with the substitution scaling rate ranging from 10^{-6} to 10^{-2}. The phylogeny of Hominidae is shown in Figure 1, and Table 1 contains information needed to generate the tree in ms.

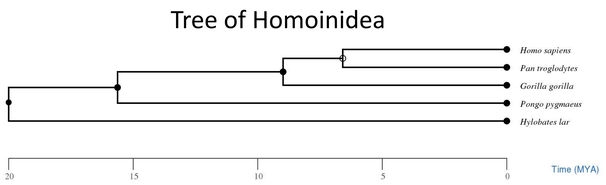

Figure 1: Simulated Tree of Homoinidea

Species	Divergence Time	Generation Time	Generations	Ne	Generations/(4*No)	Ne
Human	6.65 mya	29]	229310	10000	5.732758	7.0625
Chimp	6.65 mya	N/A	N/A	9375	N/A	7.0625
Gorilla	9.1 mya	19	478947	26250	11.97368	6.0625
Orangutans	15.8 mya	25	632000	23750	15.8	15.6250
Gibbon	20.2 mya	15	1346667	30000	33.66666	9.3125

Table 1: information need to construct tree in ms

Verification of Breakpoint

True positives and false positives were defined to be when a predicted breakpoint does and does not fall within a 250bp of a significant simulated breakpoint, respectively. Significant breakpoints are defined as those with Robinson Foulds (RF) distance > 0 or weighted Robinson Foulds distance (WRF) >= 20 [11].

Prediction of Breakpoint

- phyMl_Multi is a recombination breakpoint detection program using a hidden Markov model and the Viterbi algorithm.
- Like_HMM uses the output of phyML_Multi and performs a likelihood ratio test on the proposed recombination breakpoints using iqtree [12].
- LRTs can be used to test whether part of an alignment is better explained by a model involving one tree or two two separate trees
- All other programs were run using the default implementation within the recombination detection tool RDP4 [13]

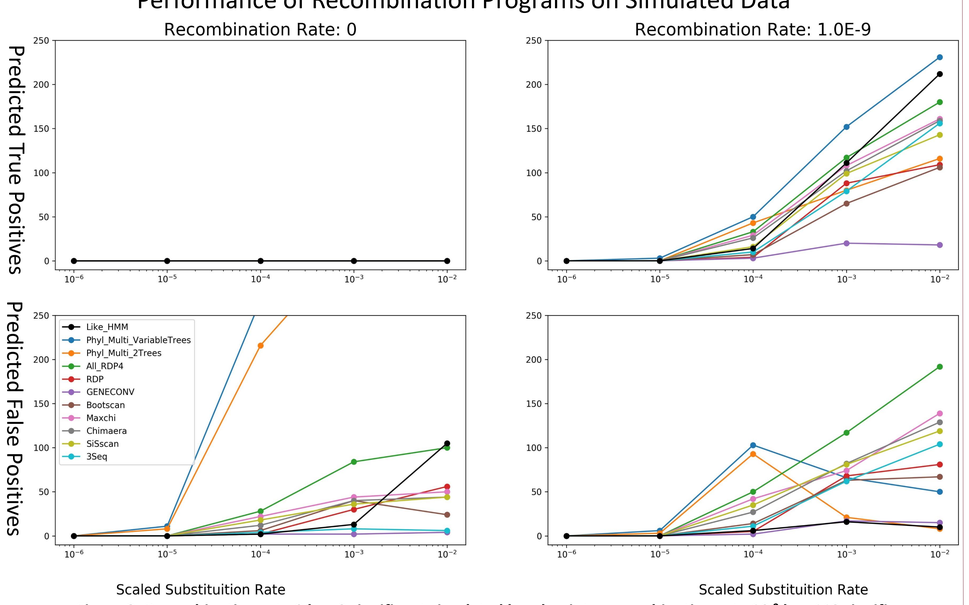

Figure 2: Recombination rate 0 has 0 significant simulated breakpoints, recombination rate 10^{-9} has 440 significant breakpoints. The first row = true positives the second row = false positives. The left column = R0 and the right column is recombination rate 10^-9.

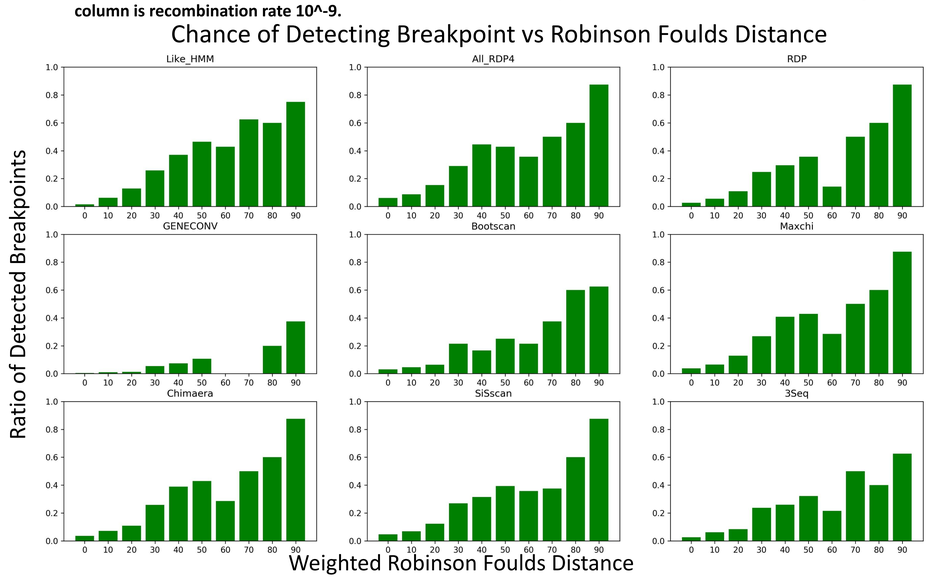

Figure 3: All simulated breakpoints are grouped into bins of size 10 based on the Weighted Robinson Foulds distance between the to alignments the breakpoint separates. The breakpoints separating alignments with large WRF values are more likely to be detected.

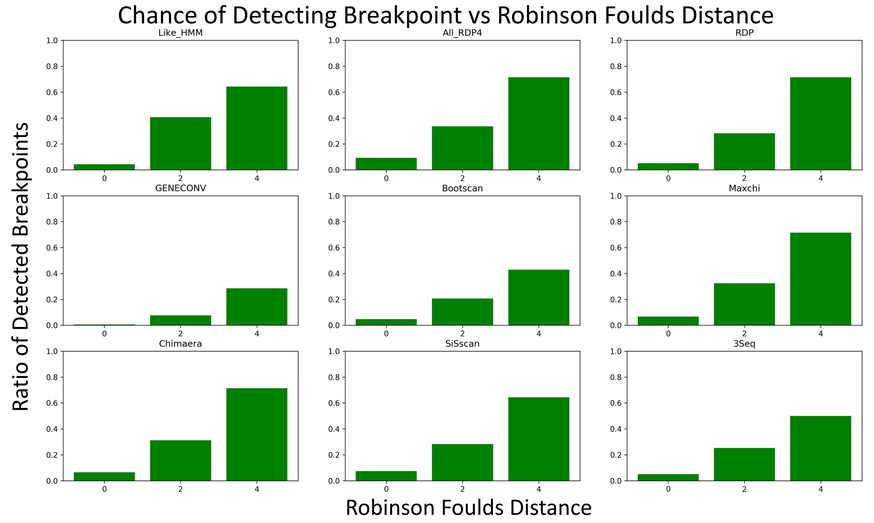

Figure 4: All simulated breakpoints are grouped based on the Robinson Foulds distance between the to alignments the breakpoint separates. The breakpoints separating alignments with large RF values are more likely to be detected.

Results

- In total there are 440 significant breakpoints for 10^-9 recombination rate. As substitution rate increases a larger percentage of these breakpoints are found.
- Most programs fail to find over half of these significant breakpoints
- Realistic substitution rates for the vertebrate tree fall between substitution scaling of 10^{-4} to 10^{-3}
- Within the realistic range Like_HMM produces fewer false positives than other methods
- Like_HMM maintains the ability to predict true recombination breakpoints
- Many simulated breakpoints do not cause topological change or have WRF less than 5. Predicting these break points will be challenging for any program due to the lack of phylogenetic signal.
- Predictive ability for all programs increases as RF distance or WRF distance increases

Conclusion

- Breakpoints which divide two distinctly different phylogenetic histories have a good chance of being located by many recombination breakpoint detection methods.
- Breakpoints dividing regions that are not distinctly different are hard to detect by all programs.
- Many existing recombination breakpoint programs produce high false positive rates when trying to pinpoint the location of recombination breakpoints. Likelihood ratio tests can successfully be used to reduce the number of false positives.

Future Work

- There are many ways to expand on the simulations done here to provide a more comprehensive overview of the performance of each recombination detection method. To start we could increase the number of taxa and include more recent recombination detection methods.
- Originally more recombination rates where simulated, but higher recombination rates divide the Homoinidea alignment into very small fragments ~100 base pairs. The accuracy of these methods could be tested on lower recombination rates.

References

[1] Posada, D., & Crandall, K. A. (2002). The effect of recombination on the accuracy of phylogeny estimation. *Journal of molecular evolution*, *54*(3), 396-402.
[2] Schierup, M. H., & Hein, J. (2000). Consequences of recombination on traditional phylogenetic analysis. *Genetics*, *156*(2), 879-891.
[3] Posada, D., & Crandall, K. A. (2002). The effect of recombination on the accuracy of phylogeny estimation. *Journal of molecular evolution*, *54*(3), 396-402.
[4] Boussau, B., Guéguen, L., & Gouy, M. (2009). A mixture model and a hidden markov model to simultaneously detect recombination breakpoints and reconstruct phylogenies. *Evolutionary Bioinformatics*, *5*, EBO-S2242.
[5] Hudson RR: Generating samples under a Wright-Fisher neutral model. Bioinformatics. 2002, 18: 337-338. 10.1093/bioinformatics/18.2.337.
[6] Kumar S, Stecher G, Suleski M, Hedges SB (2017) TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. Mol Biol Evol
[7] Prado-Martinez, J., Sudmant, P. H., Kidd, J. M., Li, H., Kelley, J. L., Lorente-Galdos, B., ... & Cagan, A. (2013). Great ape genetic diversity and population history. *Nature*, *499*(7459), 471.
[8] Chan, Y. C., Roos, C., Inoue-Murayama, M., Inoue, E., Shih, C. C., Pei, K. J. C., & Vigilant, L. (2013). Inferring the evolutionary histories of divergences in Hylobates and Nomascus gibbons through multilocus sequence data. *BMC evolutionary biology*, *13*(1), 82.
[9] Langergraber, K. E., Prüfer, K., Rowney, C., Boesch, C., Crockford, C., Fawcett, K., ... & Robbins, M. M. (2012). Generation times in wild chimpanzees and gorillas suggest earlier divergence times in great ape and human evolution. *Proceedings of the National Academy of Sciences*, *109*(39), 15716-15721.
[10] Rambaut, A., & Grass, N. C. (1997). Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Bioinformatics*, *13*(3), 235-238.
[11] Robinson, D. F., & Foulds, L. R. (1981). Comparison of phylogenetic trees. *Mathematical biosciences*, *53*(1-2), 131-147.
[12] Nguyen, L. T., Schmidt, H. A., von Haeseler, A., & Minh, B. Q. (2014). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular biology and evolution*, *32*(1), 268-274.
[13] Martin, D. P., Murrell, B., Golden, M., Khoosal, A., & Muhire, B. (2015). RDP4: Detection and analysis of recombination patterns in virus genomes. *Virus evolution*, *1*(1).