cBrother: Relaxing parental tree assumptions for Bayesian recombination detection

Fang Fang¹, Jing Ding¹, Vladimir N. Minin⁵, Marc A. Suchard⁵,⁶, and Karin S. Dorman¹,²,³*

¹Bioinformatics and Computational Biology Program, ²Department of Statistics, ³Department of Genetics, Development and Cell Biology, Iowa State University, Ames, IA 50011, USA; ⁴Ohio State University Medical Center, Columbus, OH, 43220, USA; and ⁵Department of Biomathematics and ⁶Department of Human Genetics, David Geffen School of Medicine at UCLA, Los Angeles, CA 90095, USA

ABSTRACT
Summary: Bayesian multiple change-point models accurately detect recombination in molecular sequence data. Previous Java-based implementations assume a fixed topology for the representative parental data. cBrother is a novel C language implementation that capitalizes on reduced computational time to relax the fixed tree assumption. We show that cBrother is 19 times faster than its predecessor and the fixed tree assumption can influence estimates of recombination in a medically-relevant dataset.
Availability: cBrother is freely downloadable from http://www.biomath.org/dormanks/ and can be compiled on Linux, Macintosh, and Windows operating systems. Online documentation and a tutorial are also available at the site.
Contact: kdorman@iastate.edu

INTRODUCTION
The past 20 years have yielded myriad methods for detecting rare recombination events among divergent molecular sequences. The most common methods are phylogenetic-based, inferring recombination by identifying discordant phylogenetic relationships along the sequences. The Bayesian multiple change-point (MCP) model is one such approach that simultaneously locates identifies parental genotypes and crossover-points (COPs) locations and identifies possible parental genotypes while assessing statistical support for recombination (Suchard et al., 2002). The Java package DualBrothers implements recombination detection via the Bayesian MCP (Minin et al., 2005).

To dramatically reduce the topology space and computational complexity, MCP models generally assume a fixed and known topology relates all parental genotype sequences.

Unfortunately, the fixed tree assumption fails when recombination among genotypes is possible, such as in HIV (Paraskevis et al., 2003). Even when genotype relationships are stable, only a single recombinant can be analyzed and extensive topological uncertainty within genotypes has prohibited the inclusion of multiple representative sequences per genotype. We implement a novel version of the MCP model, in C for native compilation, that relaxes the fixed parental tree assumption and uses improved likelihood calculations to substantially reduce computational run-time. cBrother both runs faster and eliminates some current restrictions of MCP models estimates recombination more accurately.

SOFTWARE DESCRIPTION
As input, cBrother takes an alignment of $N + Q$ DNA/RNA sequences in Phylip format and a command file. The first $N$ sequences are representatives for $P$ possible parental genotypes, and the last $Q$ sequences are putative recombinant sequences. Users specify the underlying evolutionary model, priors for model parameters, and Markov chain Monte Carlo (MCMC) conditions in the command file. Restarting previous chains via check-pointing is also now possible and is a useful tool for achieving MCMC convergence and crash recovery.

The user can invoke the usual fixed parental tree assumption, specify only a fixed genotype tree, or avoid all fixed tree assumptions using the command file option parent_tree. Setting parent_tree to a pre-estimated topology $\tau_N$ with $N$ terminal nodes specifies a fixed topology relating all $N$ representative sequences. Specifying instead a topology $\tau_P$ with only $P$ terminal nodes fixes just the genotype relationships. Now the set of parental trees $\{\tau_n\}$ consists of all possible $N$-taxa trees where representative sequences from the same genotype form monophyletic clades, but the branching order within genotypes varies. When the parent_tree option is set to "none," the set of parental trees is similarly constructed, except the relationship among genotypes is no
longer constrained. In all cases, the complete topology space includes all topologies produced by attaching the \( Q \) putative recombinants anywhere in tree \( \tau_N \) or all trees in \( \{ \tau_n \} \).

Experience with DualBrothers demonstrates that more than 90% of computational time is spent on likelihood calculations. Any small improvement in these calculations saves tremendous run-time. Current MCP models employ evolutionary models in which tree branch lengths are integrated out analytically. Exploiting this integration, cBrother computes and caches the finite-time transition probability matrix only once per likelihood calculation. Previous samplers recomputed this matrix along each branch and for each site of the sequence alignment.

**SPEED-UP**

We compare the run-time of cBrother to its predecessor while testing for recombination in HIV sequence L11793. The 1480bp alignment contains the putative recombinant and eight representative parental sequences. For comparison purposes, we employ both samplers to draw inference under identical models with the fixed parental tree assumption and default transition kernel options. We generate MCMC chains with 51,000 steps and discard the first 5,000 steps as burn-in. Standard diagnostics suggest adequate convergence and mixing under these conditions. cBrother takes 56sec ± 2.7sec (mean ± standard deviation, based on 10 independent runs) to simulate its chain, while DualBrothers takes 17min 53sec ± 39.8sec. Through better caching and native compilation, these results indicate that cBrother is about 19 times faster than DualBrothers. Better caching alone accounts for 15% of the improvement.

**FIXED PARENTAL TREE IMPACT**

HIV sequence U88823 is a putative genotype A1/C recombinant virus isolated from a Rwandan patient (Gao et al., 1998), but the evolutionary relationship between genotypes A1 and C varies along the genome (Anderson et al., 2000). To examine the impact of relaxing the parental tree, we consider a full-length alignment of U88823 with the consensus sequences of A1, C and three other randomly chosen genotypes. We run two independent chains under each model and check-point incrementally until stringent convergence is achieved. The final MCMC chains contain 30,000,000 steps when estimating the genotype tree and 10,000,000 when assuming a fixed genotype tree. The extra samples needed to estimate genotype trees reduce, but do not eliminate, the speed advantage of cBrother.

Both models confirm isolate U88823 is an A1/C recombinant with very high posterior probability (> 0.999). Figure 1 reports the genotype assignment to each region of U88823 along with estimated median COP locations and their posterior support. Here, COPs indicate locations where the query’s nearest neighbor changes. All COPs are well supported (posterior support > 0.95) under both models, but COP locations do not perfectly align. To quantify the difference between location estimates for the two models, we reconstruct posterior conditional location distributions for each MCMC run. These conditional distributions describe COP locations among those posterior samples that have a matching COP within a liberal range of the specified medians. The two pairs of distributions generated under the same model (fixed or relaxed parental tree) are not significantly different (\( p\)-value > 0.05 by Wilcoxon Mann-Whitney test of medians). However, distributions across models differ \( (p \ll 0.001) \), indicating that relaxing the fixed parental tree assumption can lead to significantly altered estimates. In particular, conditional distributions for the second COP (see Figure 1) are strikingly different. Accurate estimates of COP locations are necessary to understand the effects of primary and secondary sequence characteristics on promoting recombination (Galetto et al., 2004; Moumen et al., 2003). Since the difference in medians is almost twice the length of the sequence bound to reverse transcriptase when recombination occurs, an error uncertainty this large could impact downstream analyses.

**CONCLUSION**

cBrother’s improved speed, check-pointing, and ability to handle topological variation permits the analysis of larger or more complex datasets with improved accuracy. With growing numbers of recombinant sequences available, cBrother’s ability to analyze multiple recombinants will also prove useful for illuminating recombinant origins.

**ACKNOWLEDGEMENTS**

This work was supported by NIH grant GM068955.

**REFERENCES**


**Fig. 1.** Estimated recombinant structure for isolate U88823 under a fixed and relaxed parental tree. We report inferred genotypes, median COP locations, and their posterior support in brackets. Inference at the second COP is significantly altered, as shown by the location distributions obtained using a fixed (white) or relaxed (grey) tree.